

September 27, 2016

# Well-posed Bayesian Inverse Problems with Infinitely-Divisible and Heavy-Tailed Prior Measures <sup>\*</sup>

Bamdad Hosseini <sup>†</sup>

**Abstract.** We present a new class of prior measures in connection to  $\ell_p$  regularization techniques when  $p \in (0, 1)$  which is based on the generalized Gamma distribution. We show that the resulting prior measure is heavy-tailed, non-convex and infinitely divisible. Motivated by this observation we discuss the class of infinitely divisible priors and draw a connection between their tail behavior and the tail behavior of their Lévy measures. Next, we study the well-posedness of Bayesian inverse problems with heavy-tailed prior measures on Banach spaces. We establish that well-posedness relies on a balance between the growth of the log-likelihood function and the tail behavior of the prior and apply our results to special cases such as additive noise models, linear problems and infinitely divisible prior measures. Finally, we study some practical aspects of Bayesian inverse problems such as their consistent approximation and present three concrete examples of well-posed Bayesian inverse problems with infinitely-divisible prior measures.

**Key words.** Inverse problems, Bayesian, Infinitely divisible, non-Gaussian, Lévy measure.

**AMS subject classifications.** 35R30, 62F99, 60B11.

**1. Introduction.** Consider the problem of estimating a parameter  $u \in X$  from a set of measurements  $y \in Y$  where both  $X$  and  $Y$  are Banach spaces and  $y$  is associated with  $u$  through a model of the form

$$(1.1) \quad y = \tilde{\mathcal{G}}(u).$$

$\tilde{\mathcal{G}}$  is a generic stochastic mapping that models the relationship between the parameter and the observed data by taking the measurement noise into account (be it additive, multiplicative etc). As an example, if the measurement noise is additive then we can write

$$\tilde{\mathcal{G}}(u) = \mathcal{G}(u) + \eta$$

where  $\mathcal{G} : X \rightarrow Y$  is the (deterministic) *forward model* and  $\eta$  is the (random) measurement noise which is independent of  $u$ . We want to estimate the parameter  $u$  given a realization of  $y$ . Since the map  $\mathcal{G}$  may not be stably invertible this problem is in general ill-posed.

Here we consider the Bayesian framework for solution of such ill-posed problems. This approach has attracted a lot of attention in the last two decades [11, 30, 43]. The unknown parameter  $u$  is modelled as a random variable and our goal is to obtain a probability distribution on  $u$  that is informed by the data  $y$  and our prior knowledge about  $u$ . We can generate samples from this distribution and If it is concentrated around the true value of the parameter the sample mean or median will be good estimators of the true value of the parameter.

<sup>\*</sup>This work was supported in part by the Natural Sciences and Engineering Council of Canada.

<sup>†</sup>Department of mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada (bhossein@sfu.ca).

The Bayesian approach is well-established in the statistics literature [12, 3] where it is often applied in the setting where  $X, Y$  are finite-dimensional spaces. However, here we take  $X$  to be an infinite dimensional Banach space and this generalization introduces certain difficulties in our analysis.

The motivation for this level of generality is applications where the parameter  $u$  belongs to a function space such as  $L^2$  or  $BV$  (the space of functions with bounded variation). Such problems arise when the forward map involves the solution of a partial differential equation (PDE) or an integral equation such as the examples that are presented in Section 4. In practical applications we often solve these problems by discretizing the forward model and reducing the infinite dimensional problem to a finite dimensional one. However, we need to make sure that the finite dimensional posterior measure remains consistent with the infinite dimensional posterior. For example, we require that the finite dimensional posterior converges to the (true) infinite dimensional measure in the limit when the discretization is infinitely fine. Note that this notion of consistency is different from the usual definition of consistency in numerical analysis. Our concern is the convergence of probability measures rather than pointwise estimators to the solution. It turns out that ensuring this consistency is a delicate task. An example of an inconsistent discretization of an infinite dimensional inverse problem was studied in [33] where the authors demonstrated that the total variation prior loses its edge preserving properties in the limit of fine discretizations. To this end, we must study the infinite dimensional inverse problem before we construct the discrete approximation.

In this article we set out to achieve two main goals:

- G1. *Present a theory for well-posed and consistent approximation of Bayesian inverse problems with heavy-tailed prior measures.*
- G2. *Advocate the use of infinitely-divisible prior measures in modelling of prior information.*

We now discuss these goals in more detail. Let us start by introducing the infinite dimensional version of *Bayes' rule* following [43] which is understood in the sense of the Radon-Nikodym theorem [6, Thm 3.2.2]:

$$(1.2) \quad \frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z(y)} \exp(-\Phi(u; y)) \quad \text{where} \quad Z(y) = \int_X \exp(-\Phi(u; y)) d\mu_0(u).$$

Here  $\mu_0$  is the *prior measure* which reflects our prior knowledge of the parameter  $u$ ,  $\Phi(u; y)$  is the *likelihood potential* that can be thought of as the negative log of the density of the data conditioned on the parameter and  $\mu^y$  is the posterior measure on  $u$ . The posterior  $\mu^y$  is, in essence, an updated version of the prior  $\mu_0$  that is informed by the data  $y$ . The following example will clarify these notions further.

**Example 1.** Suppose  $u \in \mathbb{R}^n$  and the data  $y \in \mathbb{R}^m$  is generated via the model

$$y = \mathbf{A}u + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

where  $\mathbf{A} \in \mathbb{R}^{n \times m}$ ,  $\sigma > 0$  is fixed and  $\mathbf{I}$  is the  $m \times m$  identity matrix. Our goal is to estimate  $u$  given  $y$ . Here we are taking  $X = \mathbb{R}^n$ ,  $Y = \mathbb{R}^m$  and the forward map has the form  $\mathcal{G}(u) = \mathbf{A}u$ . Since  $\eta$  is a multivariate normal random variable we can write the likelihood potential  $\Phi(u; y)$  as:

$$\Phi(u; y) = \frac{1}{2\sigma^2} \|\mathbf{A}u - y\|_2^2.$$

Then, Bayes' rule gives

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z(y)} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{A}u - y\|_2^2\right).$$

Now suppose that

$$(1.3) \quad \frac{d\mu_0}{d\Lambda}(u) = \frac{1}{U} \exp(-\|u\|_p^p)$$

where  $d\Lambda$  denotes the Lebesgue measure on  $\mathbb{R}^n$ ,  $\|\cdot\|_p$  denotes the usual  $\ell_p$  (quasi-)norm in  $\mathbb{R}^n$  for  $p > 0$  and  $U$  is the appropriate normalizing constant. Then the posterior  $\mu^y$  is identified via its Lebesgue density as

$$(1.4) \quad \frac{d\mu^y}{d\Lambda}(u) = \frac{1}{Z(y)} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{A}u - y\|_2^2 - \|u\|_p^p\right).$$

In practice, *solving a Bayesian inverse problem* often refers to either identifying the posterior measure  $\mu^y$  (such as in (1.4)) or extracting certain statistics from it such as the mean, the variance, maximizer of the density etc. But before we can solve a Bayesian inverse problem we need to know whether the problem is well-posed to begin with (point G.1 above): Does  $\mu^y$  exist? Is it defined uniquely? Does it depend continuously on the data  $y$ ? And finally, can we approximate it in a consistent manner?

We will see later on that well-posedness depends on the behavior of the likelihood potential  $\Phi$  and the tail behavior of the prior  $\mu_0$ . We will clarify this notion with an analogy with the theory of PDEs. Consider an abstract PDE of the form

$$\mathcal{L}u = g$$

where  $\mathcal{L}$  is a differential operator and  $g$  belongs to a Banach space. We seek a solution  $u = \mathcal{L}^{-1}g$  to this problem. The well-posedness of this problem depends on the regularity of  $g$  and the smoothing behavior of  $\mathcal{L}^{-1}$ . Put simply, if  $\mathcal{L}^{-1}$  is a “nice” operator then we can solve the PDE even if  $g$  is not regular. In much the same way, we can think of a Bayesian inverse problem as that of solving the equation

$$\mathcal{P}\mu^y = \mu_0$$

where  $\mathcal{P}^{-1}$  is a mapping on the space of Radon probability measures on  $X$  that depends on the potential  $\Phi$ . Our results in Section 3 show that the behavior of  $\mathcal{P}^{-1}$  (which is identified via certain growth conditions on  $\Phi$ ) dictates the class of prior measures  $\mu_0$  that result in a well-posed inverse problem. A common theme in our results is the tradeoff between the rate of growth of  $\Phi(u; \cdot)$  and the tail behavior of  $\mu_0$ . For example, if  $\Phi(u; \cdot)$  grows like a polynomial (i.e.  $\mathcal{P}^{-1}$  is “nice”) then we can use a prior  $\mu_0$  that is heavy-tailed. On the other hand, when  $\Phi(u; \cdot)$  has exponential growth (i.e.  $\mathcal{P}^{-1}$  is not very “nice”) then the prior  $\mu_0$  should have exponentially decaying tails in order to achieve well-posedness. Then in the context of the above PDE analogy, a heavy-tailed prior is similar to a right hand side  $g$  which is not regular and the slower the growth of  $\Phi(u; \cdot)$ , the nicer the operator  $\mathcal{P}^{-1}$  becomes.

The well-posedness of Bayesian inverse problems was studied in [43, 18] for Gaussian prior measures, in [19] for Besov priors, in [28] for convex prior measures and more recently in [44] for heavy-tailed priors on separable Banach spaces. We note that our well-posedness results in this article are closely related to those of [44]. The main difference is that our theory does not rely on the assumption that the parameter space  $X$  is separable. This case is particularly interesting when one takes  $X$  to be  $C^\alpha$  (the space of Hölder continuous functions) or  $BV$ , neither of which are separable. In Section 4 we give a concrete example of an inverse problem with a  $BV$  prior that involves the edge-preserving deblurring of an image. Furthermore, we will extend the existing theory of well-posedness of Bayesian inverse problems to the case of heavy-tailed prior measures. In particular, we are interested in prior measures that are infinitely-divisible. Our interest in infinitely-divisible and heavy-tailed prior measures relies on their connection to recovery of sparse or compressible parameters. We will further clarify this connection by returning to Example 1.

**Example 1 (continued).** *The maximizer of the posterior density is referred to as the maximum a posteriori (MAP) estimate. Observe that the MAP estimate of the posterior in (1.4) is the solution to the optimization problem*

$$u_{MAP} = \arg \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2\sigma^2} \|\mathbf{A}z - y\|_2^2 + \|z\|_p^p \right\}.$$

*For  $p \geq 1$  this optimization problem is convex and can be solved efficiently. Taking  $p = 1$  results in the well-known  $\ell^1$ -regularization technique which is often used in the recovery of sparse solutions. For values of  $p \in (0, 1)$  the resulting optimization problem is no longer convex but it is known to be a good model for recovery of sparse or compressible solutions [22, 35].*

We will see that the prior distribution (1.3) for  $p \in (0, 1)$  is non-convex, heavy-tailed and infinitely divisible. Formally, a random variable  $\xi$  is infinitely-divisible if for every  $n \in \mathbb{N}$  its law coincides with the law of  $\sum_{k=1}^n \xi_k^{1/n}$  where  $\{\xi_k^{1/n}\}$  are i.i.d random variables. Thus, the above example is our first attempt at demonstrating the potential of infinitely-divisible prior measures (goal G.2). The connection between sparse recovery and heavy-tailed or infinitely divisible priors has been observed in the literature. Unser and Tafti [46] and Unser et al. [48, 47] study the sparse behavior of stochastic processes that are driven by infinitely divisible force terms and advocate their use in solution of inverse problems. More recently, Markkanen et al. [36] used a Cauchy difference prior in edge-preserving recovery of images. A detailed discussion of some heavy-tailed prior distributions such as generalizations of the student's- $t$  distribution and the  $\ell_p$ -priors can also be found in the dissertation [35]. Finally, Polson and Scott [40] and Carvalho et al. [14] propose a class of hierarchical horseshoe priors that are tailored to the recovery of sparse signals.

We now outline the structure of the remainder of this article. At the end of this section we collect some of the key definitions and notation that are used throughout the article. Section 2 is divided into two halves. In the first half we assume that the parameter space  $X$  is separable and contains an unconditional Schauder basis. Under these assumptions we will construct a few generalizations of the  $\ell_p$ -prior distribution (1.3) to measures on Banach spaces that have compressible samples. We will see that these generalized measures belong to the class of infinitely divisible probability measures and this observation motivates our exposition in the

second half of Section 2. Here, we introduce the class of infinitely-divisible prior measures and classify their tail behavior in terms of the tail behavior of their corresponding Lévy measures. In Section 3 we present some general results concerning the well-posedness of Bayesian inverse problems. The goal of this section is to identify sufficient conditions on the likelihood  $\Phi$  and the prior measure  $\mu_0$  that guarantee the well-posedness of (1.2). We dedicate Section 4 to some practical aspects of Bayesian inverse problems such as consistent approximation of the posterior measure as well as three concrete examples of well-posed inverse problems with heavy-tailed or infinitely divisible prior measures.

**1.1. Key definitions and notation.** We gather here some key definitions and assumptions. Throughout the article  $\mathbb{R}_+$  denotes the positive real line  $[0, \infty)$ . We use the shorthand notation  $a \lesssim b$  when  $a$  and  $b$  are real valued functions and there exists an independent constant  $C > 0$  such that  $a \leq Cb$ . Given two random variables  $\xi$  and  $\zeta$  we use the notation  $\xi \stackrel{d}{=} \zeta$  to denote that they have the same laws (or distributions).

We use the shorthand notation  $\{\gamma_k\}$  to denote a sequence of elements  $\{\gamma_k\}_{k=1}^\infty$  in a vector space. The usual  $\ell_p$  sequence spaces for  $p \in [1, \infty]$  are defined as the space of real valued sequences  $\{\gamma_k\}$  such that  $\|\{\gamma_k\}\|_p < \infty$  where

$$\|\{\gamma_k\}\|_p := \left( \sum_{k=1}^{\infty} |\gamma_k|^p \right)^{1/p} \quad \text{if} \quad p \in [1, \infty) \quad \text{and} \quad \|\{\gamma_k\}\|_\infty := \sup_k |\gamma_k|.$$

Similarly, we define the  $\|\cdot\|_p$  norms of finite dimensional vectors. In particular  $\|\cdot\|_2$  will denote the usual Euclidean norm. Given a positive definite matrix  $\Sigma$  of size  $m \times m$ , we define the norm

$$\|x\|_\Sigma := \|\Sigma^{-1/2}x\|_2 \quad \text{for} \quad x \in \mathbb{R}^m.$$

Throughout the article we use  $\Lambda$  to denote the Lebesgue measure in finite dimensions. Given any Borel measure  $\mu$  on  $X$  we define the spaces  $L^p(X, \mu)$  for  $p \in [1, \infty)$  as the space of  $\mu$ -equivalent classes of functions  $h : X \rightarrow \mathbb{R}$  such that  $|h|^p$  is  $\mu$ -integrable. We also use the shorthand notation  $L^p(X)$  instead of  $L^p(X, \Lambda)$  whenever we are working with the Lebesgue measure. Finally, if  $X$  is a Banach space, we use  $X^*$  to denote its topological dual.

We shall consider the prior probability measure  $\mu_0$  to be in the class of Radon probability measures on  $X$ . That is,  $\mu_0$  will be an inner regular probability measure on the Borel sets of  $X$ . Furthermore, whenever we say that  $\mu$  is a Radon probability measure on  $X$  we automatically mean that  $\mu(X) = 1$ . Finally, throughout this article we only consider complete probability measures in the following sense: If  $\mu$  is a Radon probability measure on  $X$  and  $A$  is a set of  $\mu$ -measure zero then every subset of  $A$  also has measure zero.

In this article we focus on the following notion of a well-posed Bayesian inverse problem:

**Definition 1.1 (Well-posedness).** *Suppose that  $X$  is a Banach space and  $d(\cdot, \cdot) \rightarrow \mathbb{R}$  is a metric on the space of Radon probability measures on  $X$ . Then for a choice of the prior measure  $\mu_0$  and the likelihood potential  $\Phi$ , the Bayesian inverse problem given by (1.2) is well-posed with respect to  $d$  if:*

1. (Existence and uniqueness) *There exists a unique posterior probability measure  $\mu^y \ll \mu_0$  given by Bayes' rule (1.2).*

2. (Stability) For every choice of  $\epsilon > 0$  there exists a  $\delta > 0$  so that  $d(\mu^y, \mu^{y'}) \leq \epsilon$  for all  $y, y' \in Y$  so that  $\|y - y'\|_Y \leq \delta$ .

We will study the convergence of probability measures using the Hellinger and total variation metrics on the space of probability measures on  $X$ . For two probability measures  $\mu_1$  and  $\mu_2$  that are absolutely continuous with respect to a third measure  $\nu$  on  $X$ , the total variation and Hellinger metrics are defined as

$$(1.5) \quad d_{TV}(\mu_1, \mu_2) := \frac{1}{2} \int_X \left| \frac{d\mu_1}{d\nu} - \frac{d\mu_2}{d\nu} \right| d\nu \quad \text{and} \quad d_H(\mu_1, \mu_2) := \left( \frac{1}{2} \int_X \left( \sqrt{\frac{d\mu_1}{d\nu}} - \sqrt{\frac{d\mu_2}{d\nu}} \right)^2 d\nu \right)^{1/2}.$$

Both metrics are independent of the choice of the measure  $\nu$  [6, Lem. 4.7.35]. Furthermore, convergence in one of these metrics implies convergence in the other, due to the following inequalities (see [6, Lem. 4.7.37] for a proof)

$$2d_H^2(\mu_1, \mu_2) \leq d_{TV}(\mu_1, \mu_2) \leq \sqrt{8}d_H(\mu_1, \mu_2).$$

However, one might prefer to work with the Hellinger metric as it relates directly to the error in expectation of certain functions. Suppose that  $h \in L^2(X, \mu_1) \cap L^2(X, \mu_2)$ . Then using the Radon-Nikodym theorem and Hölder's inequality one can show (see [28, Section 1] for details)

$$(1.6) \quad \left| \int_X h(u) d\mu_1(u) - \int_X h(u) d\mu_2(u) \right| \leq 2 \left( \int_X h^2(u) d\mu_1 + \int_X h^2(u) d\mu_2 \right)^{1/2} d_H(\mu_1, \mu_2).$$

For reasons that will become clear in Section 3, we prefer to study the well-posedness of inverse problems using both the Hellinger and total variation metrics. The main difference is in the restrictions that we need to impose on the prior  $\mu_0$  in order to obtain a certain rate of convergence for each metric.

**2. Infinitely divisible prior measures.** In this section we introduce the class of infinitely divisible priors. We start by presenting a generalization of the prior distribution (1.3) that was considered in Example 1. We show that this prior belongs to a larger class of distributions that are closely related to  $\ell_p$  regularization techniques. We shall extend these distributions to measures on Banach spaces with an unconditional Schauder basis and observe that they belong to the much larger class of infinitely divisible (ID) measures (see Definition 2.9). Motivated by this connection between  $\ell_p$  regularization and ID priors, we turn our attention to this class and study some of the properties of ID measures. In particular, we study the tail behavior of ID priors with respect to their Lévy measures (see Definition 2.11).

**2.1. A class of shrinkage priors with compressible samples.** When faced with the problem of recovering a sparse or compressible parameter we require the prior measure to reflect the intuition that the solution to the inverse problem is likely to have only a few large modes in some basis and the rest of the modes are negligible (see [35, Section 6.1]). Such prior distributions are often referred to as “shrinkage priors” and they have been the subject of extensive research [40, 14, 23, 16, 15]. In this section we consider a few examples of shrinkage priors that are closely related to  $\ell_p$  regularization techniques.

Most of the existing literature on properties of shrinkage priors is focused on finite dimensional problems but we will study an extension of these priors to infinite dimensional Banach spaces. Since compressibility is often considered with respect to a basis, it makes sense for us to consider a parameter space that has a basis.

Given a parameter space  $X$ , or at least a subspace  $\tilde{X} \subseteq X$  that has an unconditional Schauder basis  $\{x_k\}$ , we construct random variables of the form

$$(2.1) \quad u \sim \sum_{k=1}^{\infty} \gamma_k \xi_k x_k$$

where  $\{\gamma_k\}$  is a fixed sequence of real valued coefficients that decay sufficiently fast and the  $\{\xi_k\}$  are a sequence of independent real valued random variables that need not be identically distributed. We will take the prior measure  $\mu_0$  to be the probability measure that is induced by the random variable  $u$  in (2.1). We refer to such a prior measure  $\mu_0$  as the *product prior* obtained from  $\{\gamma_k\}$  and  $\{\xi_k\}$ . Note that this construction of the prior is reminiscent of the Karhunen-Lo  ve expansion of Gaussian measures [5, Thm. 3.5.1]. The following theorem gives sufficient conditions that ensure  $\|\cdot\|_X < \infty$   $\mu_0$ -a.s.

**Theorem 2.1.** [28, Theorem 3.9] *Suppose that  $X$  is a Banach space with an unconditional Schauder basis and let  $u$  be a random variable defined as in (2.1). If  $\{\gamma_k^2\} \in \ell_p$  and  $\{\mathbf{Var} \xi_k\} \in \ell^q$  for  $1 < p, q < \infty$  so that  $1/p + 1/q = 1$  (with  $p = 1$  for the limiting case when  $q = \infty$ ), then  $\|u\|_X < \infty$  a.s.*

**Corollary 2.2.** *In the setting of the above theorem, if the  $\{\xi_k\}$  are i.i.d.,  $\mathbf{Var} \xi_1 < \infty$  and  $\{\gamma_k\} \in \ell^2$ , then  $\|u\|_X < \infty$  a.s.*

It remains for us to show that the prior measure  $\mu_0$  that is induced by (2.1) is Radon. The assumption that  $\mu_0$  is Radon is crucial to our well-posedness results in Section 3.

**Theorem 2.3.** *Let  $\mu$  be the probability measure that is induced by the random variable  $u$  given by (2.1) where  $\{\gamma_k\}$  and  $\{\xi_k\}$  satisfy the conditions of Theorem 2.1. Then  $\mu$  is a Radon probability measure on  $X$  if the random variables  $\{\xi_k\}$  are distributed according to Radon probability measures on  $\mathbb{R}$ .*

*Proof.* Let  $\nu_k$  be the probability measure of  $\xi_k$  on  $\mathbb{R}$  and let  $A_k$  denote the support of this measure. Now define  $\mu_k$  to be the probability measure that is obtained by restricting  $\nu_k$  to  $A_k$  and consider the quasi-measure  $\tilde{\nu} = \bigotimes_{k=1}^{\infty} \mu_k$  which is generated by the countable product of the  $\mu_k$  on the product space  $A = \bigotimes_{k=1}^{\infty} A_k$ . By [7, Theorem 7.6.2]  $\tilde{\nu}$  has an extension to a Radon probability measure  $\nu$  on the product space  $A$ . Now consider the operator

$$Q : A \rightarrow X \quad Q(\{c_k\}) = \sum_{k=1}^{\infty} \gamma_k \xi_k x_k.$$

This operator is well defined since  $\|u\|_X$  is bounded on the support of  $\nu$ . It is also linear and continuous. To this end, identify  $\mu = \nu \circ Q^{-1}$ . Since the image of a Radon measure under a linear mapping is also Radon, then  $\mu$  is a Radon probability measure on  $X$ . ■

Before going further we present a result on the second raw moment of product priors which will be useful throughout the remainder of the article.



**Theorem 2.4.** Suppose that  $X$  is a Banach space with an unconditional Schauder basis  $\{x_k\}$  and let  $\mu$  be the product prior obtained from  $\{\gamma_k\} \in \ell^2$  and  $\{\xi_k\}$  where  $\xi_k$  are i.i.d and  $\mathbf{Var}\xi_k < \infty$ . Then  $\|\cdot\|_X \in L^2(X, \mu)$ .

*Proof.* Let  $u_N = \sum_{k=1}^N \gamma_k \xi_k x_k$  then for  $M > N > 0$  we have

$$\left| \int_X \|u_M\|_X^2 d\mu - \int_X \|u_N\|_X^2 d\mu \right| = \left| \int_X (\|u_M\|_X - \|u_N\|_X)(\|u_M\|_X + \|u_N\|_X) d\mu \right|$$

By Corollary 2.2 we know that  $\|u\|_X < \infty$  a.s. and so in the limit as  $M, N \rightarrow \infty$ ,  $|(\|u_M\|_X - \|u_N\|_X)| \rightarrow 2\|u\|_X$  and  $|(\|u_M\|_X - \|u_N\|_X)| \rightarrow 0$  and so  $\{\|u_N\|_X^2\}$  is Cauchy in  $L^2(X, \mu)$ . ■

We are now in position to discuss a few examples of shrinkage priors. Motivated by Example 1, we define the class of  $\ell_p$ -priors as follows:

**Definition 2.5 ( $\ell_p$ -prior).** Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$ , then we say that a Radon probability measure  $\mu$  is an  $\ell_p$ -prior on  $X$  if its samples can be expressed as  $u = \sum_{k=1}^\infty \gamma_k \xi_k x_k$  where  $\{\gamma_k\} \in \ell^2$  and  $\{\xi_k\}$  is an i.i.d sequence of real valued random variables with Lebesgue density

$$(2.2) \quad \xi_k \sim \frac{p}{2\alpha\Gamma(1/p)} \exp\left(-\frac{|t|^p}{\alpha^p}\right) d\Lambda(t)$$

where  $p \in (0, \infty)$  and  $\alpha = \sqrt{\Gamma(1/p)/\Gamma(3/p)}$ .

Here  $\Gamma$  denotes the usual Gamma function. The distribution in (2.2) belongs to the larger class of Generalized Normal distributions [38]. This class is also referred to as a Kotz-type distribution [37] or a generalized Laplace distribution [31]. Here we will not use either of these terms and simply refer to this distribution as the  $\ell_p$ -distribution to emphasize its connection to  $\ell_p$ -regularization techniques. The random variables  $\xi_k$  have bounded moments of all orders (see [38] or the discussions following the definition of the  $G_{p,q}$ -prior below), in fact

$$\mathbb{E} \xi_k^s = \frac{\alpha^s(1 + (-1)^s)}{2\Gamma(1/p)} \Gamma\left(\frac{s+1}{p}\right) \quad \text{for } s \in \mathbb{N}.$$

In particular we have that  $\mathbf{Var}\xi_k = 1$  and so it follows from Theorem 2.4 that the  $\ell_p$ -prior has bounded second moments.

Another, closely related class of priors to the  $\ell_p$ -priors can be obtained by a symmetrization of the Weibull distribution:

**Definition 2.6 ( $W_p$ -prior).** Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$ , then we say that a Radon probability measure  $\mu$  is a  $W_p$ -prior on  $X$  if its samples can be expressed as  $u = \sum_{k=1}^\infty \gamma_k \xi_k x_k$  where  $\{\gamma_k\} \in \ell^2$  and  $\{\xi_k\}$  is an i.i.d sequence of real valued random variables with Lebesgue density

$$(2.3) \quad \xi_k \sim \frac{p}{\alpha} \left(\frac{|t|}{\alpha}\right)^{p-1} \exp\left(-\frac{|t|^p}{\alpha^p}\right) d\Lambda(t),$$

where  $p \in (0, \infty)$  and  $\alpha = (2\Gamma(1 + 2/p))^{-1/2}$ .

Note that the distribution of  $\xi_k$  is simply a symmetric version of the well-known Weibull distribution [29], hence the name  $W_p$ . A straightforward calculation shows that  $\mathbf{Var}\xi_k = 1$  and once again it follows from Theorem 2.4 that the  $W_p$ -priors have bounded second moments.



Both the  $W_p$  and  $\ell_p$  distributions reduce to the Laplace distribution when  $p = 1$ . For  $p < 1$  the  $\ell_p$  distribution has non-convex level sets and puts a large portion of its mass close to the axes (see Figure 1). This behavior is amplified as  $p$  becomes small and suggests that the  $\ell_p$ -prior will incorporate sparse behavior as  $p \rightarrow 0$ .

The  $W_p$  distribution behaves very differently in comparison to the  $\ell_p$  distribution. For  $p < 1$  the  $W_p$  distribution blows up at the origin (see Figure 1(a)). This means that the  $W_p$  distribution puts more of its mass at the origin which leads us to believe that it must incorporate stronger compressibility than the  $\ell_p$  distribution.

Further insight into the behavior of the  $W_p$ -prior can be obtained by considering its MAP point estimate in finite dimensions. Formally using this prior in Example 1 gives rise to an optimization problem of the form

$$u_{\text{MAP}} = \arg \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{A}z - y\|_2^2 + \|z\|_p^p + (1-p) \sum_{k=1}^n \log(|z_k|) \right\}.$$

Of course, the log term on right hand side is not bounded from below and so we cannot gain much insight from this problem. However, we can consider a slightly modified version of this optimization problem by introducing a small parameter  $\epsilon > 0$

$$u_\epsilon = \arg \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{A}z - y\|_2^2 + \|z\|_p^p + (1-p) \sum_{k=1}^n \log(\epsilon + |z_k|) \right\}.$$

Now if  $\epsilon$  is small then the log term will heavily penalize any modes of the solution that are on a larger scale than that of  $\epsilon$  and so we expect that most of the modes of the solution  $u_\epsilon$  will be on the scale of the small parameter  $\epsilon$ . The stronger shrinkage of the posterior due to the  $W_p$ -prior is also evident in Figure 2 where we compare a prototypical example of posteriors that arise from the  $W_p$  and  $\ell_p$  priors for solution of Example 1 in 2D. Here, we clearly see that the  $W_{1/2}$ -prior results in a posterior that is highly concentrated around the axes compared to the posterior that arises from  $\ell_{1/2}$ -prior which is more spread out. Note that in either case, the posteriors are highly concentrated around the axes meaning that the map estimates as well as most of the samples from these posteriors will incorporate sparsity.

Comparing the distributions (2.2) and (2.3) suggests the definition of a larger class of priors that can interpolate between the  $\ell_p$  and  $W_p$ -priors. To this end, we introduce a new class of prior measures called the  $G_{p,q}$ -priors. The letter  $G$  is chosen due to the connection of the one dimensional version of these measures to the generalized Gamma distribution [8].

**Definition 2.7 ( $G_{p,q}$ -prior).** *Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$ , then we say that a Radon probability measure  $\mu$  is a  $G_{p,q}$ -prior on  $X$  if its samples can be expressed as  $u = \sum_{k=1}^{\infty} \gamma_k \xi_k x_k$  with  $\{\gamma_k\} \in \ell^2$  and  $\{\xi_k\}$  is an i.i.d sequence of real valued random variables with Lebesgue density*

$$(2.4) \quad \xi_1 \sim \frac{p}{2\alpha\Gamma(q/p)} \left| \frac{t}{\alpha} \right|^{q-1} \exp \left( - \left| \frac{t}{\alpha} \right|^p \right) d\Lambda(t),$$

where  $p \in (0, \infty)$  and  $\alpha = (\Gamma(q/p)/\Gamma((2+q)/p))^{1/2}$ .

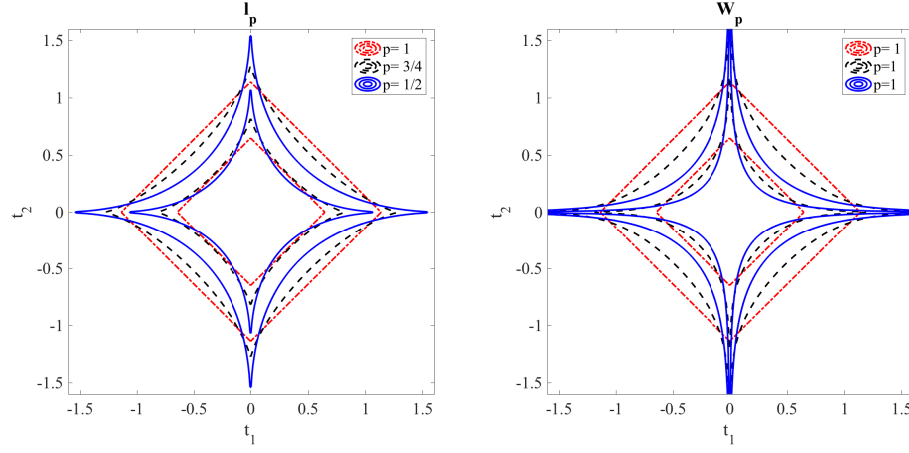


Figure 1: Contour plots of  $\ell_p$  and  $W_p$  densities in 2D for different values of  $p$ .

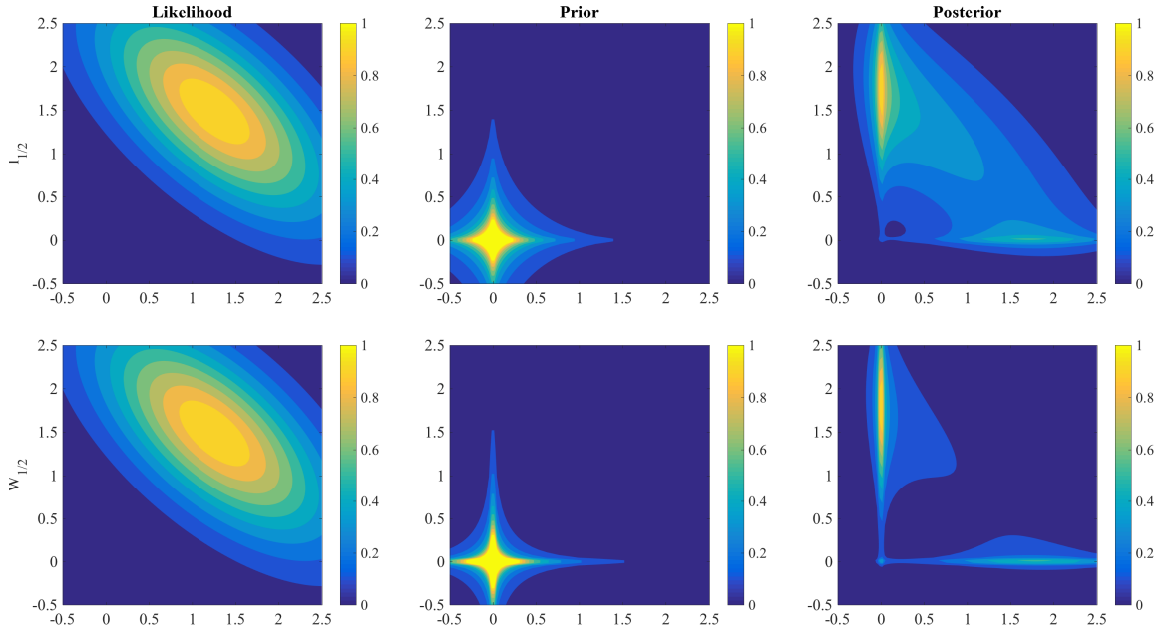


Figure 2: A prototypical example of densities that arise in the solution of Example 1 in 2D with the  $\ell_{1/2}$  (top row) and  $W_{1/2}$  priors (bottom row). From left to right columns: The likelihood that arises from the additive Gaussian noise model, the prior densities and the resulting posteriors. The densities are rescaled for better visualization.

Using the change of variables  $s = \frac{t^p}{\beta^p}$  we see that for  $k, \beta \geq 0$

$$\int_0^\infty t^k \left(\frac{t}{\beta}\right)^{q-1} \exp\left(-\left(\frac{t^p}{\beta^p}\right)\right) dt = \frac{\beta}{p} \int_0^\infty s^{\frac{k+q}{p}-1} \exp(-s) ds = \frac{\beta^{k+1}}{p} \Gamma\left(\frac{k+q}{p}\right).$$

Setting  $k = 0$  leads us to the normalizing constant in the definition of the distribution in (2.4). Furthermore, we obtain the following expression for the moments of the  $G_{p,q}$  distributions

$$\mathbb{E} |\xi_1|^s = \frac{\alpha^s (1 + (-1)^s) \Gamma((s+q)/p)}{2\Gamma(q/p)} \quad s \in \mathbb{N}.$$

In particular

$$\mathbf{Var} \xi_1 = \frac{\alpha^2 \Gamma((2+q)/p)}{\Gamma(q/p)} = 1.$$

Note that the  $\ell_p$  prior is equivalent to  $G_{p,1}$  and  $W_p$  is equivalent to  $G_{p,p}$ . Furthermore, the  $G_{1,q}$  distribution coincides with a symmetrization of the Gamma distribution. For  $q < 1$  the distribution (2.4) will blow up at the origin and so it will put a lot of its mass at zero. The  $G_{p,q}$  distributions belong to the class of ID measures by the following theorem of Bondesson.

**Theorem 2.8** ([8, Corollary 2]). *All probability density functions on  $(0, \infty)$  of the form*

$$\pi(t) = \frac{p}{\alpha \Gamma(q/p)} \left(\frac{t}{\alpha}\right)^{q-1} \exp\left(-\left(\frac{t}{\alpha}\right)^p\right)$$

are ID for  $q, \alpha > 0$  and  $0 < p \leq 1$ .

Later on we show that if the  $\xi_k$  are distributed according to an ID distribution then the corresponding product prior on  $X$  will also be an ID probability measure. Then the  $G_{p,q}$ -priors are also ID. This fact suggests the question of what other types of ID measures are good models for compressibility? We know that heavy-tailed distributions such as the Cauchy or Student's t distributions are ID and they incorporate compressible samples as well. Then there is much to be gained from the study of ID prior measures in Bayesian inverse problems. To the best of our knowledge a thorough study of the compressible behavior of ID distributions is still missing in the literature. The closest reference in this direction is the works of Unser et. al. [46, 48, 47]. While we do not study the modelling of compressible parameters, we recognize the potential impact of ID priors in this subject and so we dedicate the remainder of this section to the study of ID priors.

**2.2. Infinitely divisible priors.** We begin by collecting some results on the class of ID probability measures on Banach spaces. We only present the results that are needed in our exposition and refer the reader to [34] for a detailed introduction to ID measures on Banach spaces. Further reading can be found in the monograph [46] which contains a modern treatment of ID probability measures on nuclear spaces and the books [2, 41, 42] that are good references on the theory of ID measures in finite dimensions.

Recall that given a Borel probability measure  $\mu$  on a Banach space  $X$  its characteristic function  $\hat{\mu} : X^* \rightarrow \mathbb{C}$  is given by

$$\hat{\mu}(\varrho) = \int_X \exp(i\varrho(u)) d\mu(u) \quad \forall \varrho \in X^*.$$

Characteristic functions play a crucial role in our discussion of ID measures in this section. In what follows  $\nu^{*n}$  denotes the  $n$ -fold convolution of a measure  $\nu$  with itself.

**Definition 2.9 (ID measures [34]).** A Radon probability measure  $\mu$  on a Banach space  $X$  is called an infinitely divisible measure if for each  $n \in \mathbb{N}$  there exists a Radon probability measure  $\mu_{1/n}$  so that

$$\mu = (\mu_{1/n})^{*n}.$$

Equivalently, the probability measure  $\mu$  is ID if

$$\hat{\mu}(\varrho) = (\hat{\mu}_{1/n}(\varrho))^n \quad \forall \varrho \in X^*.$$

Put simply, a real valued random variable  $\xi$  is distributed according to an ID measure if for every  $n \in \mathbb{N}$  one can find a collection of i.i.d random variables  $\{\xi_k\}_{k=1}^n$  so that  $\xi \stackrel{d}{=} \sum_{k=1}^n \xi_k$ . Examples of such distributions include Gaussian, Laplace, Gamma, log-normal, Cauchy and Student's-t. More examples can be found in the monograph [42] where ID distributions on  $\mathbb{R}$  are studied in detail. We note that an equivalent definition of an ID measure is given as the law of a Lévy process terminated at unit time. However, we will not use this definition in order to avoid the technicalities of dealing with Lévy processes but instead we refer the interested reader to the monographs [39, 17] for further reading. The proof of the next theorem can be found in [34, Sec 5.1].

**Theorem 2.10.** Let  $\mu$  be an ID probability measure on a Banach space  $X$ . Then

- (i)  $\hat{\mu}(\varrho) \neq 0$  for all  $\varrho \in X^*$ .
- (ii) There exists a unique and continuous (in the dual norm) function  $\psi : X^* \rightarrow \mathbb{C}$  so that  $\hat{\mu}(\varrho) = \exp(\psi(\varrho))$  and  $\psi(0) = 0$ .
- (iii) If  $\mu$  is symmetric, i.e.  $\mu(A) = \mu(-A)$  for all Borel subsets  $A$  of  $X$ , then  $\hat{\mu}$  is real valued and positive.
- (iv) For every  $n \in \mathbb{N}$  the measures  $\mu_{1/n}$  are uniquely determined and  $\hat{\mu}_{1/n}(\varrho) = \exp(n^{-1}\psi(\varrho))$  for all  $\varrho \in X^*$ .

Furthermore, we define the function

$$\Psi(u, \varrho) := \exp(i\varrho(u)) - 1 - i\varrho(u)\mathbf{1}_{B_X}(u) \quad \forall u \in X, \varrho \in X^*,$$

where  $B_X$  is the unit ball in  $X$  and  $\mathbf{1}_{B_X}$  is the characteristic function of the unit ball. We recall the definition of a Lévy measure on a Banach space.

**Definition 2.11 (Lévy Measure).** A positive  $\sigma$ -finite Radon measure  $\lambda$  on  $X$  is called a Lévy measure if and only if

1.  $\lambda(\{0\}) = 0$ .
2.  $\int_X |\Psi(u, \varrho)| d\lambda(u) < \infty$  for every  $\varrho \in X^*$ .
3.  $\exp(\int_X \Psi(u, \varrho) d\lambda(u))$  is the characteristic function of a Radon probability measure on  $\mathbb{R}$  for every  $\varrho \in X^*$ .

We are now ready to present the celebrated Lévy-Khintchine representation theorem (see [34, Sec 5.7] for a proof):

**Theorem 2.12 (Lévy-Khintchine representation).** A Radon probability measure on a Banach space  $X$  is infinitely divisible if and only if there exists an element  $m \in X$ , a (positive definite)

covariance operator  $\mathcal{R} : X^* \rightarrow X$  and a Lévy measure  $\lambda$ , so that

$$(2.5) \quad \hat{\mu}(\varrho) = \exp(\psi(\varrho)) \quad \text{where} \quad \psi(\varrho) = i\varrho(m) - \frac{1}{2}\varrho(\mathcal{R}(\varrho)) + \int_X \Psi(u, \varrho) d\lambda(u).$$

Equivalently,  $\mu$  is ID precisely when there exists a point mass  $\delta_m$ , a Gaussian measure  $\mathcal{N}(0, \mathcal{R})$  and a Radon measure  $\nu$  identified via  $\hat{\nu}(\varrho) = \int_X \Psi(u, \varrho) d\lambda(u)$  so that

$$(2.6) \quad \mu = \delta_m * \mathcal{N}(0, \mathcal{R}) * \nu.$$

The Lévy-Khintchine representation implies that the triple  $(m, \mathcal{R}, \lambda)$  completely identifies an ID measure  $\mu$  and so we use the shorthand notation  $\mu = \text{ID}(m, \mathcal{R}, \lambda)$ . To gain more insight into the implications of the Lévy-Khintchine representation we recall the class of compound Poisson random variables and their corresponding probability measures.

**Definition 2.13 (Compound Poisson probability measure [34, Sec. 5.3]).** *Let  $\eta$  be a Radon probability measure on a Banach space  $X$  and suppose that  $\{u_k\}$  is a sequence of i.i.d random variables so that  $u_k \sim \eta$ . Also, let  $\tau$  be an independent Poisson random variable with rate  $c > 0$  taking values in  $\mathbb{Z}_+$ . Then  $u = \sum_{k=0}^{\tau} u_k$  is distributed according to a compound Poisson probability measure denoted by  $\text{CPois}(c, \eta)$ .*

It is straightforward to check that the characteristic function of a compound Poisson measure has the form

$$\widehat{\text{CPois}(c, \eta)}(\varrho) = \exp \left( c \int_X (\exp(i\varrho(u)) - 1) d\eta(u) \right) \quad \forall \varrho \in X^*.$$

See [34, Proposition 5.3.1] for a proof of this formula along with the fact that  $\text{CPois}(c, \eta)$  is a Radon measure on  $X$ .

Now let us return to the characteristic function of the probability measure  $\nu$  that was introduced in the Lévy-Khintchine representation (2.6)

$$(2.7) \quad \begin{aligned} \hat{\nu}(\varrho) &= \exp \left( \int_X \Psi(u, \varrho) d\lambda(u) \right) \\ &= \exp \left( \int_X (\exp(i\varrho(u)) - 1) d\lambda(u) \right) \exp \left( \int_{B_X} -i\varrho(u) d\lambda(u) \right). \end{aligned}$$

If  $0 < \lambda(X) < \infty$  then  $\lambda$  can be renormalized to define a probability measure  $\tilde{\lambda} := \frac{1}{\lambda(X)}\lambda$ . Furthermore, we can define an element  $u_\lambda \in X$  so that

$$\varrho(u_\lambda) = - \int_X \varrho(u) \mathbf{1}_{B_X}(u) d\lambda(u) \quad \forall \varrho \in X^*.$$

Putting these observations together with (2.7) gives the decomposition

$$(2.8) \quad \nu = \text{CPois}(\lambda(X), \tilde{\lambda}) * \delta_{u_\lambda}.$$

Therefore, going back to expression (2.6) we deduce that any measure  $\mu = \text{ID}(m, \mathcal{R}, \lambda)$  with  $\lambda(X) < \infty$  can be decomposed as

$$(2.9) \quad \mu = (\delta_{m+u_\lambda}) * \mathcal{N}(0, \mathcal{R}) * \text{CPois}(\lambda(X), \tilde{\lambda}).$$

In the remainder of this article we will restrict our attention to the case of ID measures with  $\lambda(X) < \infty$ . Since the tail behavior of prior measures are of importance to our well-posedness results in Section 3 we now present some results concerning the tail behavior of ID measures. We begin with the notion of a submultiplicative function.

**Definition 2.14 (Submultiplicative function).** *A non-negative, non-decreasing and locally bounded function  $h : \mathbb{R} \rightarrow \mathbb{R}^+$  is called submultiplicative if it satisfies*

$$h(t + s) \leq Ch(t)h(s) \quad \forall t, s \in \mathbb{R}$$

with an independent constant  $C > 0$

Our interest in the class of submultiplicative functions arises from the next theorem that describes some of the properties of this class.

**Theorem 2.15 ([41, Proposition 25.4]).**

- (i) *The product of two submultiplicative functions is also submultiplicative.*
- (ii) *If  $h$  is submultiplicative then so is  $(h(at + b))^\alpha$  for constants  $a, b \in \mathbb{R}$  and  $\alpha > 0$ .*
- (iii) *The functions  $\max\{1, |t|\}$  and  $\exp(|t|^\beta)$  for  $\beta \in (0, 1)$  are submultiplicative.*

*Proof.* Part (i) and (iii) follow directly from the definition and so we will only prove (ii). By the definition of submultiplicativity we have that  $h$  is non-negative and so

$$h(a(t + s) + b) = h(at + b + as + b - b) \leq C^2 h(-b)h(at + b)h(as + b).$$

Furthermore, since  $g(t) = t^\alpha$  is increasing on  $\mathbb{R}^+$  for  $\alpha > 0$  then we have that  $(h(t + s))^\alpha \leq C^\alpha (h(t))^\alpha (h(s))^\alpha$ . ■

We now present a theorem that relates the tail behavior of an ID measures to that of its Lévy measure. The following theorem was originally proved by Kruglov [32] for ID random variables on  $\mathbb{R}$  and was later generalized in [41, Theorem 25.3] to  $\mathbb{R}^n$ .

**Theorem 2.16.** *Let  $X$  be a Banach space and  $\lambda$  be a Lévy measure so that  $0 < \lambda(X) < \infty$ . Suppose that  $u \sim \mu = ID(m, \mathcal{R}, \lambda)$ ,  $\mu(X) = 1$  and  $\|\cdot\|_X < \infty$   $\mu$ -a.s. Then, given a submultiplicative function  $h$  we have that  $h(\|\cdot\|_X) \in L^1(X, \mu)$  if  $h(\|\cdot\|_X) \in L^1(X, \lambda)$ .*

*Proof.* Let  $u \sim \mu$ , then following Theorem 2.12 and the decomposition (2.9) above, we know that there exists an element  $\tilde{m} \in X$  and independent random variables  $w \sim \mathcal{N}(0, R)$  and  $v \sim \text{CPois}(\lambda(X), \frac{1}{\lambda(X)}\lambda)$  so that

$$\mathbb{E} h(\|u\|_X) = \mathbb{E} h(\|\tilde{m} + w + v\|_X) \leq C^3 h(\|\tilde{m}\|_X) (\mathbb{E} h(\|w\|_X)) (\mathbb{E} h(\|v\|_X))$$

where the inequality follows because of the triangle inequality and the fact that  $h$  is non-decreasing and locally bounded. Now by [41, Lemma 25.5] we know that there exist constants  $a, b > 0$  such that  $h(x) \leq b \exp(a|x|)$ . Using this bound with the assumption that  $\|u\|_X < \infty$   $\mu$ -a.s. along with Fernique's theorem [5, Theorem 2.8.5] for Gaussian measures on Banach spaces implies that  $\mathbb{E} h(\|w\|_X) < \infty$ . Now suppose that  $\frac{1}{\lambda(X)} \int_X h(\|u\|_X) d\lambda(u) = U < \infty$ . Then using the law of total expectation [4, Theorem 34.4], the fact that  $h$  is submultiplicative

and  $v$  is a compound Poisson random variable we obtain

$$\begin{aligned} \mathbb{E} h(\|v\|_X) &= \mathbb{E} \left( \mathbb{E} h \left( \left\| \sum_{k=0}^N v_k \right\|_X \right) \middle| N \right) \leq \mathbb{E} \left( \mathbb{E} h \left( \sum_{k=0}^N \|v_k\|_X \right) \middle| N \right) \\ &\leq \mathbb{E} \left( C^N \mathbb{E} \left( \prod_{k=0}^N h \|v_k\|_X \right) \middle| N \right) = \mathbb{E} \left( C^N \left( \prod_{k=0}^N \mathbb{E} h \|v_k\|_X \right) \middle| N \right) \\ &= \mathbb{E} ((UC)^N | N) = \sum_{k=0}^{\infty} \frac{e^{-\lambda(X)} (UC \lambda(X))^k}{k!} < \infty. \end{aligned}$$

■

By putting Theorems 2.16 and 2.15 together we immediately obtain the following corollary concerning the moments of ID measures.

**Corollary 2.17.** *Suppose that  $X$  is a Banach space and  $\mu = ID(m, \mathcal{R}, \lambda)$ . If  $\lambda$  is a Lévy measure on  $X$  so that  $0 < \lambda(X) < \infty$ ,  $\mu(X) = 1$  and  $\|\cdot\|_X < \infty$   $\mu$ -a.s. then  $\|\cdot\|_X \in L^p(X, \mu)$  whenever  $\|\cdot\|_X \in L^p(X, \lambda)$  for  $p \in [1, \infty)$ .*

*Proof.* Follows by recalling that  $\mu$  is a probability measure and  $\max\{1, |\cdot|^p\}$  is a submultiplicative function for  $p \in [1, \infty)$ . ■

Another interesting case is when the Lévy measure  $\lambda$  is convex. Recall that a Radon probability measure  $\nu$  on  $X$  is said to be convex whenever it satisfies

$$\nu(\beta A + (1 - \beta)B) \geq \nu(A)^\beta \nu(B)^{1-\beta}$$

for  $\beta \in [0, 1]$  and all Borel sets  $A$  and  $B$  (see [28, 9] for more details about convex measures). We are interested in convex measures since they have exponential tails under mild assumptions [28, Theorem 3.6]. More precisely, if  $\nu$  is a convex probability measure on  $X$  and  $\|\cdot\|_X < \infty$   $\nu$ -a.s. then there exists a constant  $0 < \kappa < \infty$  so that  $\exp(\kappa \|\cdot\|_X) \in L^1(X, \nu)$ . To this end, we have the following corollary.

**Corollary 2.18.** *Suppose that  $X$  is a Banach space and  $\mu = ID(m, \mathcal{R}, \nu)$ . If  $\nu$  is a convex probability measure on  $X$ ,  $\mu(X) = 1$  and  $\|\cdot\|_X < \infty$  a.s. under both  $\nu$  and  $\mu$  then there exists a constant  $\kappa > 0$  so that  $\exp(\kappa \|\cdot\|_X) \in L^1(X, \mu)$ .*

*Proof.* By the above discussion we have that if  $\nu$  is a convex probability measure and  $\|\cdot\|_X < \infty$   $\nu$ -a.s. then there exists a constant  $\kappa > 0$  so that  $\int_X \exp(\kappa \|\cdot\|_X) d\nu(u) < \infty$ . Since  $\exp(\kappa \|\cdot\|_X)$  is submultiplicative then the proof follows from Theorem 2.16. ■

At the end of this section we ask whether we would obtain an ID measure if we used a sequence of ID random variables in order to generate a product prior. The answer to this question is affirmative and serves as the proof of our claim that  $G_{p,q}$ -priors that were introduced earlier belong to the class of ID probability measures.

**Theorem 2.19.** *Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$  and let  $\mu$  be the product prior that is obtained from  $\{\gamma_k\} \in \ell^2$  and an i.i.d. sequence  $\{\xi_k\}$  of real valued random variables. Suppose that  $\xi_k \sim ID(0, \sigma^2, \lambda)$  where  $\sigma > 0$  and  $\lambda$  is a symmetric and finite Lévy measure on  $\mathbb{R}$  such that  $\max\{1, |\cdot|^2\} \in L^1(\mathbb{R}, \lambda)$ . Then  $\mu$  is a Radon ID*



probability measure on  $X$  with characteristic function

$$\hat{\mu}(\varrho) = \exp \left[ -\frac{1}{2} \sum_{k=1}^{\infty} \sigma^2 \gamma_k^2 \varrho(x_k)^2 + \sum_{k=1}^{\infty} \int_{\mathbb{R}} (\cos(\gamma_k \varrho(x_k) t_k) - 1) d\lambda(t_k) \right] \quad \forall \varrho \in X^*.$$

*Proof.* Since  $\max\{1, |\cdot|^2\} \in L^1(\mathbb{R}, \lambda)$ , the Lévy measure of the  $\xi_k$  has bounded second moment and so by Corollary 2.17 we see that  $\mathbf{Var} \xi_k < \infty$ . Now it follows from Corollary 2.2 that  $\|\cdot\|_X < \infty$   $\mu$ -a.s. since  $\{\gamma_k\} \in \ell^2$ . The fact that  $\mu$  is Radon follows from Theorem 2.3. Now we consider the characteristic function of  $\mu$ . By the definition of the characteristic function of a measure we have

$$\begin{aligned} \hat{\mu}(\varrho) &= \mathbb{E} \exp \left( i \varrho \left( \sum_{k=1}^{\infty} \gamma_k \xi_k x_k \right) \right) \\ &= \mathbb{E} \exp \left( \sum_{k=1}^{\infty} i \gamma_k \xi_k \varrho(x_k) \right) = \prod_{k=1}^{\infty} \mathbb{E} \exp (i \gamma_k \xi_k \varrho(x_k)) \end{aligned}$$

But the characteristic function of the  $\xi_k$  has the form

$$\hat{\xi}_k(z) = \exp \left( -\frac{1}{2} \sigma^2 z^2 + \int_{\mathbb{R}} \cos(tz - 1) d\lambda(z) \right).$$

Substituting this back into the expression for  $\hat{\mu}$  we obtain

$$\begin{aligned} \hat{\mu}(\varrho) &= \prod_{k=1}^{\infty} \exp \left( -\frac{1}{2} \sigma^2 \gamma_k^2 \varrho(x_k)^2 + \int_{\mathbb{R}} (\cos(\gamma_k \varrho(x_k) t_k) - 1) d\lambda(t_k) \right) \\ &= \exp \left[ -\frac{1}{2} \sum_{k=1}^{\infty} \sigma^2 \gamma_k^2 \varrho(x_k)^2 + \sum_{k=1}^{\infty} \int_{\mathbb{R}} (\cos(\gamma_k \varrho(x_k) t_k) - 1) d\lambda(t_k) \right] \end{aligned}$$

Now consider the sequence of measures  $\{\mu_N\}_{N=1}^{\infty}$  that are defined via their characteristic functions

$$\hat{\mu}_N(\varrho) = \exp \left[ -\frac{1}{2} \sum_{k=1}^N \sigma^2 \gamma_k^2 \varrho(x_k)^2 + \sum_{k=1}^N \int_{\mathbb{R}} (\cos(\gamma_k \varrho(x_k) t_k) - 1) d\lambda(t_k) \right].$$

Each  $\mu_N$  is ID given the fact that a finite sum of ID random variables is ID. Since the  $\{x_k\}$  are normalized and  $\{\gamma_k\} \in \ell^2$  then  $\sum_{k=1}^{\infty} \gamma_k^2 \varrho(x_k)^2 < \infty$ . Furthermore, using the inequality  $|\cos(t) - 1| \leq t^2$  we can write

$$\sum_{k=1}^{\infty} \int_{\mathbb{R}} (\cos(\gamma_k \varrho(x_k) t_k) - 1) d\lambda(t_k) \leq \sum_{k=1}^{\infty} \int_{\mathbb{R}} \gamma_k^2 \varrho(x_k)^2 t_k^2 d\lambda(t_k) = \sum_{k=1}^{\infty} \gamma_k^2 \varrho(x_k)^2 \int_{\mathbb{R}} t_k^2 d\lambda(t_k)$$

But this term is also bounded since  $\{\gamma_k\} \in \ell^2$ ,  $\{x_k\}$  are normalized and  $\max\{1, |x|^2\} \in L^1(\mathbb{R}, \lambda)$ . Then,  $\hat{\mu}_N(\ell) \rightarrow \hat{\mu}(\ell)$  for all  $\ell \in X^*$  and so the sequence  $\mu_N$  converges weakly to  $\mu$ . Therefore,  $\mu$  is also ID by Theorem [34, Theorem 5.6.2]. Observe that the Lévy measure of  $\mu$  is concentrated along the coordinate axes of the basis  $\{x_k\}$ . ■

**3. Well-posed Bayesian inverse problems.** In this section we collect certain conditions on the prior measure  $\mu_0$  and the likelihood potential  $\Phi$  that result in well-posed inverse problems. We start with minimal assumptions on the likelihood potential and forward map and make our way to more specific cases of inverse problems such as problems with linear forward maps. In a nutshell, as we put more restrictions on  $\Phi$  we are able to relax our assumptions on  $\mu_0$ . In order to help with navigation through this section we present Table 1 that collects our main results and the key underlying assumptions.

Theorem/Corollary	Main assumptions	type of result
Thm. 3.2	$\Phi$ is locally bounded and Lipschitz in $u$ .	$\mu^y$ is well-defined
Thm. 3.3	$\Phi$ satisfies Assumption 1	$\mu^y$ depends continuously on $y$
Cor. 3.5	$\Phi$ has polynomial growth in $u$ and $\mu_0$ has finitely many moments	well-posedness
Cor. 3.7	$\Phi \geq 0$ in addition to Assumption 1	well-posedness
Thm. 3.9	$Y = \mathbb{R}^m$ , forward map is linear and bounded, measurement noise is additive and Gaussian	well-posedness
Cor. 3.13	$Y = \mathbb{R}^m$ , forward map is linear and bounded, measurement noise is additive and Gaussian, $G_{p,q}$ -prior	well-posedness
Cor. 3.14	$Y = \mathbb{R}^m$ , forward map is linear and bounded, measurement noise is additive and Gaussian, prior is ID	well-posedness

Table 1: Summary of the key theorems and corollaries of Section 3. In each case we identify the key underlying assumptions as well as the type of final result.

We begin by identifying some conditions on  $\Phi$  that allow us to use a very large class of prior measures including those that are heavy-tailed.

**Assumption 1.** Suppose that  $X$  and  $Y$  are Banach spaces and the likelihood potential  $\Phi : X \times Y \rightarrow \mathbb{R}$  satisfies the following properties:

- (i) (Lower bound in  $u$ ): There is a positive and non-decreasing function  $f_1 : \mathbb{R}_+ \rightarrow [1, \infty)$  so that  $\forall r > 0$ , there is a constant  $M(r) \in \mathbb{R}$  such that  $\forall u \in X$  and  $\forall y \in Y$  with  $\|y\|_Y < r$ ,

$$\Phi(u; y) \geq M - \log(f_1(\|u\|_X)).$$

- (ii) (Boundedness above):  $\forall r > 0$  there is a constant  $K(r) > 0$  such that  $\forall u \in X$  and  $\forall y \in Y$  with  $\max\{\|u\|_X, \|y\|_Y\} < r$ ,

$$\Phi(u; y) \leq K.$$

- (iii) (Continuity in  $u$ ):  $\forall r > 0$  there exists a constant  $L(r) > 0$  such that  $\forall u_1, u_2 \in X$  and  $y \in Y$  with  $\max\{\|u_1\|_X, \|u_2\|_X, \|y\|_Y\} < r$ ,

$$|\Phi(u_1; y) - \Phi(u_2; y)| \leq L\|u_1 - u_2\|_X.$$

- (iv) (Continuity in  $y$ ): There is a positive and non-decreasing function  $f_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  so that  $\forall r > 0$ , there is a constant  $C(r) \in \mathbb{R}$  such that  $\forall y_1, y_2 \in Y$  with  $\max\{\|y_1\|_Y, \|y_2\|_Y\} < r$

and  $\forall u \in X$ ,

$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq C f_2(\|u\|_X) \|y_1 - y_2\|_Y.$$

We reiterate that an analog of our well-posedness results in this section can be found in the work of Sullivan [44] for the case when  $X$  is a separable Banach space. In fact, it was shown in [44] that if  $X$  is a separable Banach space then one can achieve well-posedness in the Hellinger metric if Assumption 1 (i) and (iii) are replaced by the assumption that  $\Phi(u; y)$  is a Caratheodory function [1, Section 4.10]. Here, we require the stronger Lipschitz continuity assumption in order to prove the existence of the posterior when  $X$  is not separable (see Example 3 in Section 4).

Our first task is to establish the existence and uniqueness of the posterior measure. The following result on the concentration of Radon probability measures on Banach spaces is central to our argument.

**Theorem 3.1** ([7, Thm. 7.12.4]). *Let  $\mu$  be a Radon probability measure on a Banach space  $X$ . Then, there exists a reflexive and separable Banach space  $E$  embedded in  $X$  such that  $\mu(X \setminus E) = 0$  and the closed balls of  $E$  are compact in  $X$ .*

**Theorem 3.2.** *Suppose  $X$  is a Banach space,  $\mu_0$  is a Radon probability measure on  $X$  and let  $\Phi$  satisfy Assumptions 1 (i), (ii) and (iii) with a function  $f_1 \geq 1$ . If  $f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$  then the posterior  $\mu^y$  given by (1.2) is a well-defined Radon probability measure on  $X$ .*

*Proof.* Our proof will closely follow the approach of [43, Thm. 4.1] and [28, Thm. 2.2]. Assumption 1(iii) implies the continuity of  $\Phi$  on  $X$  which in turn implies that  $\Phi(\cdot, y) : X \rightarrow \mathbb{R}$  is  $\mu_0$ -measurable. We will now show that the normalizing constant satisfies  $0 < Z(y) < \infty$  which proves that  $\mu^y$  is well-defined. The fact that  $\mu^y$  is Radon will then follow from the absolute continuity of  $\mu^y$  with respect to  $\mu_0$  and the assumption that  $\mu_0$  is Radon as well [7, Lem. 7.1.11].

Following Assumption 1(i) we can write

$$\begin{aligned} Z(y) &= \int_X \exp(-\Phi(u; y)) d\mu_0(u) \\ &\leq \int_X \exp(\log(f_1(\|u\|_X)) - M) d\mu_0(u) \\ &= \exp(-M) \int_X f_1(\|u\|_X) d\mu_0(u) < \infty. \end{aligned}$$

We now need to show that the normalizing constant  $Z(y)$  does not vanish. Let  $E \subset X$  be the embedded Banach space of Theorem 3.1 with the norm  $\|\cdot\|_E$ . For a constant  $R > 0$  we can find  $R' > 0$  so that  $\{\|u\|_E \leq R\} \subseteq \{\|u\|_X < R'\}$ . Then it follows from Assumption 1 that for  $R > 0$

$$\begin{aligned} Z(y) &= \int_X \exp(-\Phi(u; y)) d\mu_0(u) \\ &\geq \int_{\{\|u\|_X < R'\}} \exp(-K) d\mu_0(u) \\ &\geq \int_{\{\|u\|_E \leq R\}} \exp(-K) d\mu_0(u) = \exp(-K) \mu_0(\{\|u\|_E \leq R\}). \end{aligned}$$

Recall that  $\mu_0$  is supported on  $E$  and the closed balls of  $E$  are compact in  $X$ . Furthermore,  $\mu_0(\{\|u\|_E < R\}) > 0$  for large enough  $R$ . This is true because if the measure of centred balls of  $E$  are zero then the measure of all balls of  $E$  would have to be zero. Since  $E$  is separable, it can be covered by a countable union of balls which would imply  $\mu_0(E) = 0$ . This contradicts the fact that  $\mu_0$  is a probability measure concentrated on  $E$ . ■

We now establish the stability of Bayesian inverse problems with respect to perturbations in the data. Similar versions of the following theorems are available for Gaussian priors in [43], for Besov priors in [19], for convex priors in [28] and for heavy-tailed priors on separable Banach spaces in [44]. We will establish the stability of the inverse problems with respect to both the total variation and the Hellinger metrics. The reason for this choice is that one needs slightly stronger assumptions on the prior measure in order to achieve the same rate of convergence in the Hellinger metric as compared to total variation.

**Theorem 3.3.** *Suppose that  $X$  is a Banach space,  $\mu_0$  is a Radon probability measure on  $X$  and  $\Phi$  satisfies Assumptions 1(i), (ii) and (iv) with functions  $f_1, f_2$ . Let  $\mu^y$  and  $\mu^{y'}$  be two measures defined via (1.2) for  $y$  and  $y' \in Y$ , both absolutely continuous with respect to  $\mu_0$ . If  $f_2(\|\cdot\|_X)f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$  then  $\forall r > 0$ , there exists a constant  $C(r) > 0$  with  $\max\{\|y\|_Y, \|y'\|_Y\} < r$  so that  $d_{TV}(\mu^y, \mu^{y'}) \leq C\|y - y'\|_Y$ .*

*Proof.* Consider the normalizing constants  $Z(y)$  and  $Z(y')$ . We have already established in the proof of Theorem 3.2 that neither of these constants will vanish and they are both bounded. Thus the measures  $\mu^y$  and  $\mu^{y'}$  are well-defined. Applying the mean value theorem to the exponential function and using Assumptions 1(i), (iv) and the assumption that  $f_2(\|\cdot\|_X)f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$  we obtain

$$\begin{aligned}
 |Z(y) - Z(y')| &\leq \int_X \exp(-\Phi(u; y)) |\Phi(u; y) - \Phi(u; y')| d\mu_0(u) \\
 &\leq \left( \int_X \exp(\log(f_1(\|u\|_X)) - M) C f_2(\|u\|_X) d\mu_0(u) \right) \|y - y'\|_Y \\
 &\leq C \exp(-M) \left( \int_X f_1(\|u\|_X) f_2(\|u\|_X) d\mu_0(u) \right) \|y - y'\|_Y \\
 &\lesssim \|y - y'\|_Y.
 \end{aligned}
 \tag{3.1}$$

Following the definition of the total variation distance we have

$$\begin{aligned}
 2d_{TV}(\mu^y, \mu^{y'}) &= \int_X |Z(y)^{-1} \exp(-\Phi(u; y)) - Z(y')^{-1} \exp(-\Phi(u; y'))| d\mu_0(u) \\
 &\leq \int_X |Z(y)^{-1} \exp(-\Phi(u; y)) - Z(y')^{-1} \exp(-\Phi(u; y))| d\mu_0(u) \\
 &\quad + Z(y')^{-1} \int_X |\exp(-\Phi(u; y)) - \exp(-\Phi(u; y'))| d\mu_0(u) \\
 &=: I_1 + I_2.
 \end{aligned}$$

Now using (3.1) we have

$$I_1 = |Z(y)^{-1} - Z(y')^{-1}| Z(y) = \frac{Z(y)}{Z(y')} |Z(y') - Z(y)| \lesssim \|y - y'\|_X.$$

Furthermore, using the mean value theorem, Assumption 1 (i) and (iv) we can write

$$\begin{aligned}
Z(y')I_2 &= \int_X |\exp(-\Phi(u; y)) - \exp(-\Phi(u; y'))| d\mu_0(u) \\
&\leq \int_X \exp(-\Phi(u; y)) |\Phi(u; y') - \Phi(u; y)| d\mu_0(u) \\
&\leq C \exp(-M) \left( \int_X \exp(\log(f_1(\|u\|_X))) f_2(\|u\|_X) d\mu_0(u) \right) \|y - y'\|_Y \\
&\lesssim \|y - y'\|_Y.
\end{aligned}$$

■

We can obtain a similar result for the Hellinger metric but we require a stronger assumption concerning the integrability of  $f_1$  and  $f_2$  functions.

**Theorem 3.4.** *Consider the setting of Theorem 3.3 with the stronger assumption that  $(f_2(\|\cdot\|_X))^2 f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$ . Then  $\forall r > 0$  there exists a constant  $C(r) > 0$  with  $\max\{\|y\|_Y, \|y'\|_Y\} < r$  so that  $d_H(\mu^y, \mu^{y'}) \leq C\|y - y'\|_Y$ .*

**Proof.** Following the definition of the Hellinger metric and using the inequality  $(a + b)^2 \leq 2(a^2 + b^2)$ , we can write

$$\begin{aligned}
2d_H^2(\mu^y, \mu^{y'}) &= \int_X \left( Z(y)^{-1/2} \exp\left(-\frac{1}{2}\Phi(u; y)\right) - Z(y')^{-1/2} \exp\left(-\frac{1}{2}\Phi(u; y')\right) \right)^2 d\mu_0(u) \\
&\leq \frac{2}{Z(y)} \int_X \left( \exp\left(-\frac{1}{2}\Phi(u; y)\right) - \exp\left(-\frac{1}{2}\Phi(u; y')\right) \right)^2 d\mu_0(u) \\
&\quad + 2 \left| Z(y)^{-1/2} - Z(y')^{-1/2} \right|^2 \int_X \exp(-\Phi(u; y')) d\mu_0(u). \\
&=: I_1 + I_2.
\end{aligned}$$

Once again using the mean value theorem and Assumptions 1(iv) and (i) we get

$$\begin{aligned}
\frac{Z(y)}{2} I_1 &\leq \int_X \frac{1}{4} \exp(-\Phi(u; y)) |\Phi(u; y') - \Phi(u; y)|^2 d\mu_0(u) \\
&\leq \int_X \frac{1}{4} \exp(-\Phi(u; y)) C^2 (f_2(\|u\|_X))^2 \|y - y'\|_Y^2 d\mu_0(u) \\
&\leq \frac{C^2}{4} \int_X \exp(\log(f_1(\|u\|_X)) - M) (f_2(\|u\|_X))^2 \|y - y'\|_Y^2 d\mu_0(u) \\
&= \frac{C^2}{4} \exp(-M) \left( \int_X f_1(\|u\|_X) (f_2(\|u\|_X))^2 d\mu_0(u) \right) \|y - y'\|_Y^2 \lesssim \|y - y'\|_Y^2.
\end{aligned}$$

Furthermore, since (3.1) still holds then we have

$$I_2 = 2|Z(y)^{-1/2} - Z(y')^{-1/2}|^2 Z' \lesssim |Z(y) - Z(y')|^2 \lesssim \|y - y'\|_Y^2.$$

■

Theorems 3.3 and 3.4 are very similar. The main distinction is that in order to obtain the same rate of convergence in the Hellinger metric we need a (possibly) stronger assumption

regarding the integrability of  $f_1(\|u\|_X)$  and  $f_2(\|u\|_X)$ . So far we encountered conditions of the form  $(f_2(\|u\|_X))^p f_1(\|u\|_X) \in L^1(X, \mu_0)$  for  $p \in \{0, 1, 2\}$ . Intuitively, these conditions identify the interplay between the growth of  $\Phi(u; y)$  as a function of  $\|u\|_X$  and the tail behavior of the prior  $\mu_0$ . This connection becomes clear in the next two corollaries.

**Corollary 3.5.** *Suppose that  $\Phi$  satisfies the conditions of Assumption 1 with  $f_1(t) = f_2(t) = \max\{1, |t|^p\}$  for  $p \geq 0$  and  $\mu_0$  is a Radon probability measure on  $X$ . If  $\max\{1, \|\cdot\|_X^{2p}\} \in L^1(X, \mu_0)$  then the Bayesian inverse problem (1.2) is well-posed in both the total variation and Hellinger metrics.*

**Corollary 3.6.** *Suppose that  $\mu_0$  is a Radon probability measure on  $X$  and  $\exp(\kappa\|\cdot\|_X) \in L^1(X, \mu_0)$  for some constant  $\kappa > 0$ . Then the Bayesian inverse problem (1.2) is well-posed in both the total variation and Hellinger metrics whenever  $\Phi$  satisfies the conditions of Assumption 1 with functions  $f_1, f_2$  that are polynomially bounded.*

For the remainder of this section we will focus on specific classes of likelihood potentials  $\Phi$  which allow us to further relax our assumption regarding the tail behavior of  $\mu_0$ . We start with the case of additive noise models and consider linear inverse problems afterwards. Our results on both of these problems are viewed as specific cases of the results above.

**3.1. The case of additive noise models.** Additive noise models have a special place in practical applications due to their convenience and flexibility [30]. Therefore, we dedicate this section to Bayesian inverse problems that arise from this type of measurement model. Suppose that the data is finite dimensional and, without loss of generality, take  $Y = \mathbb{R}^m$ ,  $m \in \mathbb{N}$ . Now suppose that  $y \in Y$  is related to the parameter  $u \in X$  via the model

$$(3.2) \quad y = \mathcal{G}(u) + \eta \quad \eta \sim \pi(y) d\Lambda(y)$$

where  $\pi(y)$  is the Lebesgue density of the measurement noise  $\eta$  and  $\mathcal{G} : X \rightarrow \mathbb{R}^m$  is the forward map. It is straightforward to check that under these assumptions

$$(3.3) \quad \Phi(u; y) = -\log \pi(\mathcal{G}(u) - y).$$

In particular if  $\eta \sim \mathcal{N}(0, \Sigma)$  with an  $m \times m$  positive definite matrix  $\Sigma$  then

$$(3.4) \quad \Phi(u; y) = \frac{1}{2} \|(\mathcal{G}(u) - y)\|_{\Sigma}^2.$$

Now if  $\log \pi(y) \leq 0$  (which is clearly the case when  $\eta$  is Gaussian or Laplace) then  $\Phi(u; y)$  will satisfy Assumption 1(i) with the constant  $M = 0$  and  $f_1(x) = 1$ . This observation will allow us to relax our assumption on the tail behavior of the prior whenever the measurement noise is additive.

**Corollary 3.7.** *Suppose that  $X$  is a Banach space and  $\Phi(u; y) \geq 0$  and it satisfies Assumptions (ii) and (iv) with a function  $f_2$ . Suppose that the prior measure  $\mu_0$  is a Radon probability measure on  $X$  and let  $\mu^y$  and  $\mu^{y'}$  be two measures defined via (1.2) for  $y$  and  $y' \in \mathbb{R}^m$ . Then the posterior measure  $\mu^y$  is well-defined and*

- (i) *If  $f_2(\|\cdot\|_X) \in L^1(X, \mu_0)$  then  $\forall r > 0$  with  $\max\{\|y\|_Y, \|y'\|_Y\} < r$  there exists a constant  $C(r) > 0$  so that  $d_{TV}(\mu^y, \mu^{y'}) \leq C\|y - y'\|_Y$ .*
- (ii) *If  $f_2(\|\cdot\|_X) \in L^2(X, \mu_0)$  then  $d_H(\mu^y, \mu^{y'}) \leq C\|y - y'\|_Y$ .*

At this point it is natural to identify conditions on the distribution of the noise and the forward operator that guarantee that the likelihood potential of (3.3) satisfies the conditions of Assumption 1. We will address this when  $\eta$  is Gaussian but our approach can be generalized to other types of additive noise models.

**Theorem 3.8.** *Consider the additive noise model of (3.2) when  $\eta \sim \mathcal{N}(0, \Sigma)$  and  $\Sigma$  is a positive-definite matrix. Then the corresponding likelihood potential  $\Phi(u; y) \geq 0$ . Furthermore,  $\Phi$  satisfies the conditions of Assumption 1(iv) with  $f_2(x) = 1 + \tilde{f}(x)$  if there is a positive, non-decreasing and locally bounded function  $\tilde{f} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  so that the following conditions are satisfied:*

(i) *There is a constant  $C > 0$  for which*

$$\|\mathcal{G}(u)\|_2 \leq C\tilde{f}(\|u\|_X) \quad \forall u \in X.$$

(ii)  *$\forall r > 0$  there is a constant  $K(r) > 0$  so that for all  $u_1, u_2 \in X$  and  $\max\{\|u_1\|_X, \|u_2\|_X\} < r$*

$$\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_2 \leq K\|u_1 - u_2\|_X.$$

*Proof.* Since we assumed that  $\eta$  is Gaussian then the likelihood potential is of the form (3.4). Then it is clear that  $\Phi(u; y) \geq 0$  which immediately implies that  $\Phi$  satisfies Assumption 1(i) with  $M = 0$  and  $f_1(x) = 0$ . Now fix  $r > 0$  and suppose that  $u \in X$  and  $y \in Y$  so that  $\max\{\|u\|_X, \|y\|_Y\} < r$ . Define  $\tilde{r} = \max\{r, C\tilde{f}(r)\}$  and note that  $\tilde{r}$  is bounded since we assumed that  $\tilde{f}$  is locally bounded. To this end

$$\Phi(u; y) \leq \|\mathcal{G}(u)\|_\Sigma^2 + \|y\|_\Sigma^2 \lesssim \tilde{r}^2.$$

Thus  $\Phi$  satisfies Assumption 1(ii).

Now we will show that  $\Phi$  will satisfy Assumption 1(iii) as well. Let  $r$  and  $\tilde{r}$  be defined as above and consider  $u_1, u_2 \in X$  and  $y \in Y$  so that  $\max\{\|u_1\|_X, \|u_2\|_X, \|y\|_Y\} < r$ . Then using the identity  $\|a\|_2^2 - \|b\|_2^2 = \langle a - b, a + b \rangle$  for  $a, b \in \mathbb{R}^m$  and the conditions (i) and (ii) of the theorem we obtain

$$\begin{aligned} 2|\Phi(u_1; y) - \Phi(u_2; y)| &= \left| \|\mathcal{G}(u_1) - y\|_\Sigma^2 - \|\mathcal{G}(u_2) - y\|_\Sigma^2 \right| \\ &= \left| \langle \Sigma^{-1/2}(\mathcal{G}(u_1) - \mathcal{G}(u_2)), \Sigma^{-1/2}(\mathcal{G}(u_1) + \mathcal{G}(u_2) - 2y) \rangle \right| \\ &\leq (\|\mathcal{G}(u_1)\|_\Sigma + \|\mathcal{G}(u_2)\|_\Sigma + 2\|y\|_\Sigma) \|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_\Sigma \\ &\leq C(\tilde{r})\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_\Sigma \leq 2K(r)\|u_1 - u_2\|_X. \end{aligned}$$

Finally, fix  $r > 0$  and consider  $y_1, y_2 \in Y$  so that  $\max\{\|y_1\|_2, \|y_2\|_2\} < r$ . Then for any  $u \in X$ , using a similar approach as above we can write

$$\begin{aligned} 2|\Phi(u; y_1) - \Phi(u; y_2)| &= \left| \|\mathcal{G}(u) - y_1\|_\Sigma^2 - \|\mathcal{G}(u) - y_2\|_\Sigma^2 \right| \\ &= \left| \langle \Sigma^{-1/2}(y_2 - y_1), \Sigma^{-1/2}(2\mathcal{G}(u) - y_1 - y_2) \rangle \right| \\ &\leq (\|y_2\|_\Sigma - \|y_1\|_\Sigma + 2\|\mathcal{G}(u)\|_\Sigma) \|y_2 - y_1\|_\Sigma \\ &\leq C(r)(1 + \tilde{f}(\|u\|_X))\|y_1 - y_2\|_2. \end{aligned}$$

■



**3.2. The case of linear inverse problems.** Now we study the well-posedness of a smaller class of inverse problems. We assume that the likelihood potential has the form

$$\Phi(u; y) : X \times \mathbb{R}^m \rightarrow \mathbb{R}, \quad \Phi(u; y) = \frac{1}{2} \|\mathcal{G}(u) - y\|_{\Sigma}^2$$

where  $\Sigma$  is a positive definite matrix and  $\mathcal{G} : X \rightarrow \mathbb{R}^m$  is bounded and linear. This case is of particular importance to us due to its occurrence in the Compressed Sensing literature [22] and estimation of sparse parameters. In this case, we can further relax our conditions on the prior measure  $\mu_0$  and achieve well-posedness so long as the prior  $\mu_0$  has bounded first moment.

**Theorem 3.9.** *Let  $X$  be a Banach space and  $Y = \mathbb{R}^m$ . Suppose that the forward map  $\mathcal{G} : X \rightarrow \mathbb{R}^m$  is bounded and linear and consider the additive noise model*

$$y = \mathcal{G}(u) + \eta \quad \text{where} \quad \eta \sim \mathcal{N}(0, \Sigma) \quad \text{and} \quad \Sigma \text{ is positive definite.}$$

*Then the Bayesian inverse problem of identifying the posterior  $\mu^y$  via (1.2) is well-posed in both the Hellinger and total variation metrics if the prior  $\mu_0$  is a Radon probability measure on  $X$  and  $\|\cdot\|_X \in L^1(X, \mu_0)$ .*

*Proof.* Since  $\mathcal{G}$  is bounded and linear then there is a constant  $C > 0$  such that

$$\|\mathcal{G}(u)\|_2 \leq C\|u\|_X, \quad \|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_2 \leq C\|u_1 - u_2\|_X.$$

This implies that  $\mathcal{G}$  satisfies conditions (i) and (ii) of Theorem 3.8 with  $\tilde{f}(x) = x$ . Then by Theorem 3.8 we have that the resulting quadratic likelihood potential  $\Phi(u; y)$  of the inverse problem satisfies Assumption 1(iv) with  $f_2(x) = 1 + x$ . Putting this together with Corollary 3.7(i) and Theorem 3.2, we can see that our inverse problem is well-posed for any prior measure  $\mu_0$  for which  $\|\cdot\|_X \in L^1(X, \mu_0)$ . ■

At the end of this section we return to the product priors of Section 2.1 and show that those priors result in well-posed Bayesian inverse problems under the linear and additive noise assumptions.

**Theorem 3.10.** *Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$  and take  $Y = \mathbb{R}^m$ . Suppose that the measurement noise is additive and Gaussian and the forward map  $\mathcal{G}$  is bounded and linear. Furthermore, suppose that  $\mu_0$  is a product prior with sample paths  $u = \sum_{k=1}^{\infty} \gamma_k \xi_k x_k$  where  $\{\gamma_k\} \in \ell^2$  and  $\{\xi_k\}$  are i.i.d and  $\mathbf{Var} \xi_k < \infty$ . Then the inverse problem (1.2) is well-posed in both the total variation and Hellinger metrics.*

*Proof.* The fact that  $\mu_0$  is a Radon probability measures on  $X$  follows from Theorem 2.3 and Corollary 2.2. Now if  $\mathbf{Var} \xi_k < \infty$  then  $\mathbb{E} \xi_k^2 < \infty$  as well and so it follows from Theorem 2.4 that  $\|\cdot\|_X \in L^2(X, \mu)$ . Then the assertion follows from Theorems 3.8 and 3.7. ■

We now consider the the  $\ell_p$ ,  $W_p$  and  $G_{p,q}$ -priors of Section 2. The proof of the following Corollaries follow directly from Theorem 2.4 and the fact that the  $G_{p,q}$  distributions in 1D have bounded variance for  $p, q > 0$ .

**Corollary 3.11.** *Let  $X$  be a Banach space with an unconditional Schauder basis  $\{x_k\}$  and  $Y = \mathbb{R}^m$ . Suppose that the measurement noise is additive and Gaussian and that the forward map  $\mathcal{G}$  is bounded and linear. Then the Bayesian inverse problem (1.2) is well-posed in both the Hellinger and total variation metrics if  $\mu_0$  is an  $\ell_p$ -prior with  $p > 0$ .*

**Corollary 3.12.** *The result of Corollary 3.11 holds if  $\mu_0$  is a  $W_p$ -prior with  $p > 0$ .*

**Corollary 3.13.** *The result of Corollary 3.11 holds if  $\mu_0$  is a  $G_{p,q}$ -prior with  $p, q > 0$ .*

Finally, we consider the case of ID priors.

**Corollary 3.14.** *Let  $X$  be a Banach space and  $Y = \mathbb{R}^m$ . Consider the additive noise model:*

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \Sigma)$$

where  $\Sigma$  is a positive definite matrix and  $\mathcal{G} : X \rightarrow \mathbb{R}^m$  is bounded and linear. Also, suppose that  $\mu_0 = ID(m, \mathcal{R}, \lambda)$  where  $\lambda$  is a Lévy measure such that  $\lambda(X) < \infty$ ,  $\mu_0(X) = 1$  and  $\|\cdot\|_X < \infty$   $\mu_0$ -a.s. Then the Bayesian inverse problem (1.2) is well-posed if  $\|\cdot\|_X \in L^1(X, \lambda)$ .

*Proof.* The assertion is a direct consequence of Theorems 2.17 and 3.9. ■

**Corollary 3.15.** *Consider the setting of Theorem 3.14 except that  $\mathcal{G}$  is not linear and instead it satisfies the conditions of Theorem 3.8 with a submultiplicative function  $f$ . Then the Bayesian inverse problem (1.2) is well-posed if  $1 + \tilde{f}(\|\cdot\|_X) \in L^1(X, \lambda)$ .*

**4. Practical considerations and examples.** We now turn our attention to practical aspects of solving an inverse problem within the Bayesian framework. In the first part of this section we discuss the problem of approximating the posterior measure via approximation of the likelihood potential. Afterwards, we will present three concrete examples of Bayesian inverse problems with heavy-tailed priors that arise from practical problems in image deblurring and ultrasound therapy.

**4.1. Consistent approximation of the posterior.** Up to this point we were concerned with identifying prior measures  $\mu_0$  that result in a well-posed Bayesian inverse problem for a given likelihood potential  $\Phi$ . However, in practice we cannot solve the inverse problem directly on the infinite dimensional Banach space. Therefore, we need to obtain a finite dimensional approximation to the posterior measure  $\mu^y$  which is, in some sense, consistent with the infinite dimensional limit.

To this end, we will define the notion of *consistent approximation* of a Bayesian inverse problem in the context of applications where one would discretize (1.2) by approximating the likelihood potential  $\Phi$  with a discretized version  $\Phi_N$ , akin to a finite element discretization. We define the approximation  $\mu_N^y$  to  $\mu^y$  via

$$(4.1) \quad \frac{d\mu_N^y}{d\mu_0} = \frac{1}{Z_N(y)} \exp(-\Phi_N(u; y)) \quad \text{where} \quad Z_N(y) = \int_X \exp(-\Phi_N(u; y)) d\mu_0(u).$$

**Definition 4.1 (Consistent approximation[28]).** *The approximate Bayesian inverse problem (4.1) is a consistent approximation to (1.2) for a choice of  $\mu_0$ ,  $\Phi$  and  $\Phi_N$  if  $d(\mu^y, \mu_N^y) \rightarrow 0$  as  $|\Phi(u; y) - \Phi_N(u; y)| \rightarrow 0$ . Here,  $d$  is either the total variation or the Hellinger metric.*

Note that this notion of a consistent approximation relates directly to practical applications. Suppose, for example, that we are interested in computing the expected value of a quantity  $h(u)$  under the posterior  $\mu^y$  but we can only compute the expectation under the approximation  $\mu_N^y$ . Then if  $\mu_N^y$  is a consistent approximation in the Hellinger metric then we have, by the bound (1.6), that if  $h \in L^2(X, \mu^y) \cap L^2(X, \mu_N^y)$  then

$$\left| \int_X h(u) d\mu^y(u) - \int_X h(u) d\mu_N^y(u) \right| \leq C d_H(\mu^y, \mu_N^y).$$

Thus the error in computing the expected value of  $h$  will vanish as  $N \rightarrow \infty$ . In what follows, we will provide sharper bounds on the rate of convergence of the distances between  $\mu^y$  and  $\mu_N^y$  under mild conditions.

**Theorem 4.2.** *Assume that the measures  $\mu^y$  and  $\mu_N^y$  are defined via (1.2) and (4.1), for a fixed  $y \in Y$  and all values of  $N$ , and are absolutely continuous with respect to the prior  $\mu_0$  which is a Radon probability measure on  $X$ . Also assume that both  $\Phi$  and  $\Phi_N$  satisfy Assumptions 1(i) and (ii) with an appropriate function  $f_1$ , uniformly for all  $N$ . Furthermore, assume that there exists a positive and non-decreasing function  $f_3 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  so that*

$$(4.2) \quad |\Phi(u; y) - \Phi_N(u; y)| \leq f_3(\|u\|_X) \rho(N)$$

where  $\rho(N) \rightarrow 0$  as  $N \rightarrow \infty$ .

(i) *If  $f_3(\|\cdot\|_X) f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$  then there exists a constant  $D > 0$ , independent of  $N$  such that*

$$d_{TV}(\mu^y, \mu_N^y) \leq D \rho(N).$$

(ii) *If  $(f_3(\|\cdot\|_X))^2 f_1(\|\cdot\|_X) \in L^1(X, \mu_0)$  then there exists a constant  $D' > 0$ , independent of  $N$  such that*

$$d_H(\mu^y, \mu_N^y) \leq D' \rho(N).$$

**Proof.** Our method of proof is very similar to that of Theorems 3.3 and 3.4 and so we will only present it for the total variation distance. The existence and uniqueness of the measures  $\mu^y$  and  $\mu_N^y$  follows from Theorem 3.2 for all values of  $N$ . Next, using the mean value theorem followed by Assumption 1(i), (4.2) and the assumption that  $f_3(\|\cdot\|_X) f_1(\|\cdot\|_X)$  is  $\mu_0$ -integrable gives

$$\begin{aligned} |Z(y) - Z_N(y)| &\leq \int_X \exp(-\Phi(u; y)) |\Phi(u; y) - \Phi_N(u; y)| d\mu_0(u) \\ &\leq \left( \int_X \exp(\log(f_1(\|u\|_X)) - M) C f_3(\|u\|_X) d\mu_0(u) \right) \rho(N) \\ &\lesssim \rho(N). \end{aligned}$$

Furthermore, we have

$$\begin{aligned} 2d_{TV}(\mu^y, \mu^{y'}) &= \int_X |Z(y)^{-1} \exp(-\Phi(u; y)) - Z_N(y)^{-1} \exp(-\Phi_N(u; y))| d\mu_0(u) \\ &\leq \int_X |Z(y)^{-1} \exp(-\Phi(u; y)) - Z_N(y)^{-1} \exp(-\Phi(u; y))| d\mu_0(u) \\ &\quad + Z_N(y)^{-1} \int_X |\exp(-\Phi(u; y)) - \exp(-\Phi_N(u; y))| d\mu_0(u) \\ &=: I_1 + I_2. \end{aligned}$$

It then follows in a similar manner to proof of Theorem 3.3 that  $I_1 \lesssim \rho(N)$  and  $I_2 \lesssim \rho(N)$  which gives the desired result. ■

We now consider a more specific example of consistent approximations when the prior measure  $\mu_0$  has a product structure. Suppose that the likelihood potential  $\Phi$  satisfies the

Assumption 1 with some functions  $f_1, f_2$ . Also, assume that the space  $X$  has an unconditional Schauder basis  $\{x_k\}$ . Now consider the sequence of spaces  $(X_N, \|\cdot\|_X)$  where  $X_N = \text{span}\{x_k\}_{k=1}^N$ . These are linear subspaces of  $X$  and for each  $N \in \mathbb{N}$  we have  $X = X_N \oplus X_N^\perp$ , meaning that every  $u \in X$  can be written as  $u = u_N + u_N^\perp$  where  $u_N \in X_N$  and  $u_N^\perp \in X_N^\perp$ .

Suppose that the prior measure  $\mu_0$  has the product structure of (2.1) and assume that it has sufficiently fast decaying tails so that the posterior measure  $\mu^y$  is well-defined. Observe that for every value of  $N$  the product prior can be factored as

$$(4.3) \quad \mu_0 = \mu_N \otimes \mu_N^\perp$$

where  $\mu_N$  and  $\mu_N^\perp$  are Radon measures on  $X_N$  and  $X_N^\perp$ . It is natural for us to discretize the potential  $\Phi$  using a projection operator:

$$(4.4) \quad \Phi_N(u; y) := \Phi(P_N u; y)$$

where  $P_N : X \rightarrow X_N$  is defined by  $P_N(u) = u_N$ . Next, define the approximate posterior measures  $\mu_N^y$  as in (4.1) using the above definition of  $\Phi_N$ . Observe that, under these assumptions, the  $\mu_N^y$  will factor as (see [28, Section 4.1] for the details of why this is valid)

$$(4.5) \quad \mu_N^y = \nu_N \otimes \mu_N^\perp.$$

where

$$\frac{d\nu_N}{d\mu_N} = \frac{1}{Z_N(y)} \exp(-\Phi(P_N u; y)).$$

In other words, the likelihood potential is only informative on the subspace  $X_N$  and so by comparing (4.3) and (4.5) we see that the approximate posterior  $\mu_N^y$  differs from the prior only on this subspace and it is identical to the prior on  $X_N^\perp$ . As an example, we now check whether this method for discretization of the posterior results in a consistent approximation to  $\mu^y$  in the additive Gaussian noise case.

**Theorem 4.3.** *Consider the above setting where the posterior and the prior have the prescribed product structures and the  $X_N$  are linear subspaces of  $X$ . Suppose that  $\Phi$  and  $\Phi_N$  are given by*

$$\Phi(u; y) = \frac{1}{2} \|\mathcal{G}(u) - y\|_2^2, \quad \Phi_N(u; y) = \frac{1}{2} \|\mathcal{G}(P_N u) - y\|_2^2$$

where  $P_N : X \rightarrow X_N$  is the projection operator that was defined before. Assume that the following conditions are satisfied:

- (a)  $\|u - P_N u\|_X \leq \|u\|_X \rho(N)$ .
- (b)  $\|\mathcal{G}(u)\|_2 \leq C \tilde{f}_1(\|u\|_X) \quad \forall u \in X$ .
- (c)  $\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_2 \leq \tilde{f}_2(\max\{\|u_1\|_X, \|u_2\|_X\}) \|u_1 - u_2\|_X \quad \forall u_1, u_2 \in X$ .

Here  $\rho$  is a positive function such that  $\rho(N) \rightarrow 0$  as  $N \rightarrow \infty$  and the functions  $\tilde{f}_1, \tilde{f}_2 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  are non-decreasing and locally bounded and  $\tilde{f}_1 \geq 1$ . Then

- (i) If  $\tilde{f}_1(\|\cdot\|_X) \tilde{f}_2(\|\cdot\|_X) \in L^1(X, \mu_0)$  then there exists a constant  $D > 0$  independent of  $N$  so that

$$d_{TV}(\mu^y, \mu_N^y) \leq D\rho(N).$$

(ii) If  $\tilde{f}_1(\|\cdot\|_X)\tilde{f}_2(\|\cdot\|_X) \in L^2(X, \mu_0)$  then there exists a constant  $D' > 0$  independent of  $N$  so that

$$d_H(\mu^y, \mu_N^y) \leq D'\rho(N).$$

*Proof.* It follows from Theorem 3.8 that  $\Phi$  and  $\Phi_N$  satisfy Assumption 1 uniformly in  $N$  with  $M = 0$ ,  $f_1(x) = 1$ . Then the measures  $\mu^y$  and  $\mu_N^y$  are well-defined for all values of  $N \in \mathbb{N}$  by Theorem 3.2. Now it follows from our assumptions on  $\mathcal{G}$  that

$$\begin{aligned} 2|\Phi(u; y) - \Phi_N(u; y)| &= \left| \|(\mathcal{G}(u) - y)\|_{\Sigma}^2 - \|(\mathcal{G}(P_N u) - y)\|_{\Sigma}^2 \right| \\ &= \left| \langle \Sigma^{-1/2}(\mathcal{G}(u) - \mathcal{G}(P_N u)), \Sigma^{-1/2}(\mathcal{G}(u) + \mathcal{G}(P_N u) - 2y) \rangle \right| \\ &\leq (\|\mathcal{G}(u)\|_{\Sigma} + \|\mathcal{G}(P_N u)\|_{\Sigma} + 2\|y\|_{\Sigma}^2) \|(\mathcal{G}(u) - \mathcal{G}(P_N u))\|_{\Sigma} \\ &\leq C\tilde{f}_1(\|u\|_X)\tilde{f}_2(\|u\|_X)\|u - P_N u\|_X. \end{aligned}$$

The claim will now follow by taking  $f_3(x) = \tilde{f}_1(x)\tilde{f}_2(x)$  and applying Theorem 4.2. ■

A few comments are in order concerning the previous theorem. First, the function  $\rho(N)$  is independent of the forward map and the prior and depends solely on the topology of  $X$ . Then the rate of convergence of  $\mu_N^y$  to  $\mu^y$  depends directly on the rate of convergence of  $P_N$  to the identity map in the operator norm. Also, observe that in order to achieve the same rate of convergence in the Hellinger metric as in the total variation metric, we need to impose stronger tail assumptions on the prior  $\mu_0$ . Finally, we have the following corollary concerning the setting where  $\mathcal{G}$  is bounded and linear:

**Corollary 4.4.** *Consider the setting of Theorem 4.3 with the exception that  $\mathcal{G} : X \rightarrow \mathbb{R}^m$  is bounded and linear. If  $\|\cdot\|_X \in L^2(X, \mu_0)$  then there exists a constant  $D$  independent of  $N$  so that*

$$d_H(\mu^y, \mu_N^y) \leq D\rho(N).$$

This corollary along with the identity (1.6) imply that for the case of linear problems with additive Gaussian noise, we only need  $\mu_0$  to have bounded moments of order two in order to be able to control the error in computing the expectation of functions  $h \in L^2(X, \mu^y) \cap L^2(X, \mu_N^y)$ . In fact we have the inequality

$$\left| \int_X h(u) d\mu^y(u) - \int_X h(u) d\mu_N^y(u) \right| \leq C\rho(N).$$

**4.2. Example 2: Deconvolution.** We now turn our attention to a few concrete examples of inverse problems with heavy-tailed or non-Gaussian prior measures. Our first example concerns the deconvolution problem which is a classic example of a linear inverse problem with wide applications in optics and imaging [49, 24]. This problem was also considered in [28] as an example problem with a convex prior measure.

Let  $X = L^2(\mathbb{T})$  where  $\mathbb{T}$  is the circle of radius  $(2\pi)^{-1}$  and let  $Y = \mathbb{R}^m$  for a fixed integer  $m$ . Suppose that  $\eta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$  where  $\sigma \in \mathbb{R}$  is a fixed constant and  $\mathbf{I}$  is the  $m \times m$  identity matrix. Let  $S : C(\mathbb{T}) \rightarrow \mathbb{R}^m$  be a bounded linear operator that collects point values of a continuous function on a set of  $m$  points over  $\mathbb{T}$ . Finally, given a fixed kernel  $\varphi \in C^1(\mathbb{T})$ , define the forward map  $\mathcal{G} : X \rightarrow Y$  as

$$(4.6) \quad \mathcal{G}(u) = S(\varphi * u) \quad \text{where} \quad (\varphi * u)(t) := \int_{\mathbb{T}} \varphi(t-s)u(s) d\Lambda(s).$$

Now suppose that the data  $y$  is generated via  $y = \mathcal{G}(u) + \eta$  and our goal is to estimate the original image  $u$  given noisy point values of its blurred version. Note that our assumptions so far imply a quadratic likelihood potential of the form (3.4)

It follows from Young's inequality [26, Thm. 13.8] that  $(\varphi * \cdot) : L^2(\mathbb{T}) \rightarrow L^2(\mathbb{T})$  is a bounded linear operator and furthermore,  $(\varphi * u) \in C^1(\mathbb{T})$  for all  $u \in L^2(\mathbb{T})$ . Since pointwise evaluation is a bounded linear functional on  $C^1(\mathbb{T})$  then the forward map  $\mathcal{G} : L^2(\mathbb{T}) \rightarrow \mathbb{R}^m$  is bounded and linear and so we can use the results of Section 3.2 to show the well-posedness of this inverse problem.

We now construct our prior measure using the product priors of Section 2.1. Consider the functions

$$\tilde{w}(t) = \begin{cases} 1 & 0 \leq t \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad \tilde{v}(t) = \begin{cases} 1 & 0 \leq t \leq 1/2, \\ 1 & 1/2 \leq t \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function  $\tilde{v}$  is the Haar wavelet and  $\tilde{w}$  is its corresponding scaling function. Following [20, Section 9.3], we can define the periodic functions

$$w_{jn}(t) := \sum_{l \in \mathbb{Z}} \phi(2^j(t+l) - n), \quad v_{jn}(t) := \sum_{l \in \mathbb{Z}} \psi(2^j(t+l) - n),$$

as well as the functions

$$x_1(t) = w_{0,0}(t), \quad x_2(t) = v_{0,0}(t), \quad x_{2^j+n_j+1}(t) = v_{j,n_j}(t).$$

for  $j \in \mathbb{Z}_+$  and  $n_j = \{0, 1, \dots, 2^j - 1\}$ . The  $\{x_k\}$  form an orthonormal basis for  $L^2(\mathbb{T})$  and so they can be used in the construction of a  $G_{p,q}$ -prior.

We now choose  $p, q \in (0, 1)$  and take the prior  $\mu_0$  to be the  $G_{p,q}$ -prior obtained using the wavelet basis  $\{x_k\}$  and the fixed sequence  $\{\gamma_k\}$  where

$$\gamma_{2^j+n_j+1} = (1 + |2^{j+1}|^2)^{-1/2} \quad \forall n \in \mathbb{Z}_+.$$

Clearly,  $\{\gamma_k\} \in \ell^2$  and so it follows from Theorem 2.2 that  $\|u\|_{L^2(\mathbb{T})} < \infty$  a.s. Furthermore, we know that the  $G_{p,q}$ -priors have bounded moments of order two. Putting this together with the fact that the forward map  $\mathcal{G}$  is bounded and linear we immediately obtain the well-posedness of this inverse problem using Theorem 3.10.

**4.3. Example 3: Deconvolution with a BV prior.** In Example 2 above we constructed our prior measure on the separable space  $L^2(\mathbb{T})$ . However, throughout the article we mainly considered the setting where the parameter space  $X$  was not separable. Here we present a concrete example that utilizes a prior that is supported on a non-separable Banach space. This example is inspired by the work of Markkanen et. al [36] who used a Cauchy difference prior for edge-preserving deconvolution of images. Edge-preserving recovery is a well-known inverse problem and it is often solved by total variation regularization techniques [49]. In [33] the authors showed that using a total variation prior in the Bayesian framework results in an inconsistent Bayesian inverse problem in the sense that in the limit as the discretization of the forward problem becomes very fine, the total variation prior loses its edge preserving

property and converges to a Gaussian prior. Here, we will formulate a consistent version of this problem using a  $BV$  prior.

We define the space of functions of bounded variation on the circle  $\mathbb{T}$  as

$$BV(\mathbb{T}) := \{u \in L^1(\mathbb{T}) : \|u\|_{BV} := \|u\|_{L^1(\mathbb{T})} + \|\partial u\|_{(C^1(\mathbb{T}))^*} < \infty\}.$$

Here  $\partial u$  is understood as an element in the algebraic dual of  $C^1(\mathbb{T})$ . Equipped with the above norm,  $BV(\mathbb{T})$  is a Banach space but it is not separable [10].

Now consider the deconvolution problem that was described in Example 2 above but take  $X = BV(\mathbb{T})$ . Since the convolution kernel  $\varphi \in C^1(\mathbb{T})$  and  $BV(\mathbb{T}) \subset L^1(\mathbb{T})$  then the forward map  $\mathcal{G} : BV(\mathbb{T}) \rightarrow \mathbb{R}^m$  (given by (4.6)) is well-defined, bounded and linear. Thus the likelihood potential has the form (3.4) once more.

We will now construct our prior measure by considering a periodization of a Lévy process on the interval  $[0, 1)$ . Let  $u(t)$  for  $t \in [0, 1)$  denote a stochastic process such that

$$u(0) = 0, \quad \hat{u}_t(s) = \mathbb{E} \exp(isu(t)) = \exp(t\psi(s)) \quad s \in \mathbb{R}.$$

Here  $\hat{u}_t(s)$  is the characteristic function of  $u(t)$ . We assume that the function  $\psi$  has the form

$$\psi(s) = \int_{\mathbb{R}} \exp(i\xi s) - 1 - i\xi s \mathbf{1}_{[-1,1]} d\nu(\xi),$$

where the measure  $\nu$  is a Lévy measure that satisfies

$$\int_{\{|\xi| \leq 1\}} |\xi| d\nu(\xi) < \infty.$$

Note that the function  $\psi$  is the characteristic function of a pure jump Lévy process, i.e. a Lévy process without the Brownian motion component. This assumption implies that the sample paths of  $u(t)$  have countably many jump discontinuities and so  $\|u(t)\|_{BV([0,t])} < \infty$  a.s. [17, Proposition 3.9]. As an example, one can take  $\nu = \mathcal{N}(0, 1)$  which implies that  $u(t)$  is a compound Poisson process with piecewise constant sample paths and normal jumps.

We now take the prior measure  $\mu_0$  to be the probability measure that is induced by the periodic versions of  $u(t)$ . With an abuse of notation we use  $u$  to denote the corresponding periodic process on  $\mathbb{T}$ . Since  $\|u\|_{BV(\mathbb{T})} < \infty$ ,  $\mu_0$ -a.s. and  $\mathcal{G} : BV(\mathbb{T}) \rightarrow \mathbb{R}^m$  is bounded and linear we immediately obtain the well-posedness of this inverse problem via Theorem 3.9.

**4.4. Example 4: Quadratic measurements of a continuous field.** As our final example, we will consider a problem with a non-linear forward map. Our goal is to estimate a continuous field from quadratic measurements of its point values. This inverse problem was encountered in [27] in recovery of aberrations in high intensity focused ultrasound treatment and it is closely related to the phase retrieval problem [21, 25, 13]. Let  $X = C(\mathbb{T})$  and let  $\{t_k\}_{k=1}^n$  be a collection of distinct points in  $\mathbb{T}$ . Now define the operator

$$S : C(\mathbb{T}) \rightarrow \mathbb{R}^n \quad (S(u))_j = u(t_j) \quad j = 1, 2, \dots, n.$$

This operator collects point values of functions in  $C(\mathbb{T})$ . Let  $\{z_k\}_{k=1}^m$  be a fixed collection of vectors  $z_k \in \mathbb{R}^n$  and define the forward map

$$\mathcal{G} : C(\mathbb{T}) \rightarrow \mathbb{R}^m, \quad (\mathcal{G}(u))_j := |z_j^T S(u)|^2 \quad \text{for } j = 1, 2, \dots, m,$$



which collects quadratic measurements of the point values of a continuous function. We complete our model of the measurements with an additive layer of Gaussian noise

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}),$$

where  $\sigma > 0$ . Our goal in this problem is to infer the function  $u \in C(\mathbb{T})$  from the quadratic measurements  $y$ .

Our first task is to identify the behavior of the forward map. A straightforward calculation shows that

$$(4.7) \quad \|\mathcal{G}(u)\|_2 \leq \tilde{K} \|S(u)\|_2^2 \leq K \|u\|_{C(\mathbb{T})}^2,$$

where  $\tilde{K}, K > 0$  are constants that are independent of  $u$  but depend on the  $z_k$ . Note that the final inequality follows because pointwise evaluation is a bounded linear operator on  $C(\mathbb{T})$ .

Furthermore, we have that for  $u_1, u_2 \in C(\mathbb{T})$

$$\begin{aligned} (\mathcal{G}(u_1) - \mathcal{G}(u_2))_j &= |z_j^T S(u_1)|^2 - |z_j^T S(u_2)|^2 \\ &= (z_j^T (S(u_1) - S(u_2))) (z_j^T (S(u_1) + S(u_2))) \\ &\leq D_j (\max\{\|u_1\|_{C(\mathbb{T})}, \|u_2\|_{C(\mathbb{T})}\}) \|u_1 - u_2\|_{C(\mathbb{T})}. \end{aligned}$$

Here, the constant  $D_j > 0$  depends on  $z_j$ . We can now use this bound in order to obtain

$$(4.8) \quad \|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_2 \leq D (\max\{\|u_1\|_{C(\mathbb{T})}, \|u_2\|_{C(\mathbb{T})}\}) \|u_1 - u_2\|_{C(\mathbb{T})}$$

where the constant  $D > 0$  will only depend on the  $D_j$ . Observe that the above bounds in (4.7) and (4.8) imply that  $\mathcal{G}$  satisfies the conditions of Theorem 3.8 with a function  $\tilde{f}(x) = x^2$ . Therefore, that theorem implies that the likelihood function  $\Phi$  for our problem will satisfy Assumption 1 (iv) with  $f_2(x) = 1 + x^2$ . Now we use Corollary 3.7 to infer that well-posedness can be achieved if we choose a prior measure  $\mu_0$  for which  $f_2(\|\cdot\|) = 1 + \|\cdot\|_{C(\mathbb{T})}^2 \in L^1(C(\mathbb{T}), \mu_0)$ .

In order to construct such a prior measure  $\mu_0$  we will consider a product prior with samples of the form

$$u \sim \sum_{k \in \mathbb{Z}} \gamma_k \xi_k w_k \quad \text{where} \quad w_k(t) = (2\pi)^{-1/2} \exp(2\pi i k t).$$

Note that the  $\{w_k\}$  are simply the Fourier basis on  $\mathbb{T}$ . Our plan is to construct the prior measure so that it is supported on a sufficiently regular Sobolev space and then use the Sobolev embedding theorem to show that the prior samples belong to  $C(\mathbb{T})$ . The reason for going through the Sobolev space is the fact that  $C(\mathbb{T})$  does not have an unconditional Schauder basis and so we cannot directly apply the methodology of Section 2.1.

To this end, we choose

$$\gamma_k = (1 + |k|^2)^{-3/2} \quad k \in \mathbb{Z},$$

and suppose that the  $\{\xi_k\}$  are i.i.d and  $\xi_1 \sim \text{CPois}(0, \text{Lap}(0, 1))$  (recall Definition 2.13), where  $\text{Lap}(0, 1)$  is the standard Laplace distribution on the real line with Lebesgue density  $\pi(x) = \frac{1}{2} \exp(-|x|)$  which clearly has exponential tails and this, in turn, implies that  $\text{Var} \xi_1 < \infty$ . Note that the random variables  $\xi_k$  have a positive probability of being zero and hence draws

from this prior will incorporate a certain level of sparsity. Observe that this is a different type of sparsity in comparison to the  $G_{p,q}$ -prior. Samples from this compound Poisson prior have a non-zero probability of having modes that are exactly zero. The samples from the  $G_{p,q}$ -prior have a zero probability of having modes that are exactly zero and instead most of their modes will concentrate in a neighborhood of zero.

Recall that the Sobolev space  $H^1(\mathbb{T})$  is defined as

$$H^1(\mathbb{T}) := \{v \in L^2(\mathbb{T}) : \|v\|_{H^1(\mathbb{T})}^2 := \sum_{k \in \mathbb{Z}} (1 + |k|^2) |\langle v, w_k \rangle|^2 < \infty\}$$

where  $\langle \cdot, \cdot \rangle$  is the usual  $L^2(\mathbb{T})$  inner product. Now consider  $u \sim \mu_0$  then

$$\|u\|_{H^1(\mathbb{T})}^2 = \sum_{k \in \mathbb{Z}} (1 + |k|^2)^{-1} |\xi_k|^2.$$

But  $\{(1 + |k|^2)^{-1}\} \in \ell^1$  and  $\mathbf{Var}|\xi_k|^2 < \infty$  and so it follows from Corollary 2.2 that  $\|u\|_{H^1(\mathbb{T})}^2 < \infty$  a.s. Now the Sobolev embedding theorem [45, Proposition 3.3] guarantees that  $\|u\|_{C(\mathbb{T})} < \infty$  a.s. and it follows from Theorem 2.4 that  $\|u\|_{C(\mathbb{T})}^2 \in L^1(C(\mathbb{T}), \mu_0)$ .

**5. Closing remarks.** We began this article by introducing two main goals: Present a theory of well-posedness for Bayesian inverse problems that includes heavy-tailed and ID priors (goal *G.1*) and motivate the study of ID prior measures and demonstrate their potential in modelling of prior information (goal *G.2*). We started by focusing on the latter goal and introduced the  $\ell_p$ ,  $W_p$  and  $G_{p,q}$ -prior measures. We showed that this class is closely related to  $\ell_p$  regularization techniques in sparse recovery and shows great potential for practical applications. We observed that the  $G_{p,q}$ -priors are ID and this motivated our study of the ID class. We introduced the class of ID prior measures and used the celebrated Lévy-Khintchine theorem to show that the tail behavior of an ID measure is tied to the tail behavior of its Lévy measure. Our discussions in this direction motivated the need for extending the theory of well-posed Bayesian inverse problems to the case of heavy-tailed prior measures.

Our approach to well-posedness theory was to identify the minimal restrictions on the prior measure given a choice of the likelihood potential  $\Phi$ . A common theme in our results was the trade-off between the tail decay of the prior and the growth of the likelihood potential. As an example, we considered the setting where the likelihood had a quadratic form and the forward map was linear. This example corresponds to linear inverse problems with additive Gaussian noise that are of great interest in practice. We showed that in this simple setting well-posedness can be achieved if the prior has moments of order one.

Finally, we considered some practical aspects of solving inverse problems with heavy-tailed or ID priors. We discussed consistent discretization of inverse problems and the use of projections in discretization of the likelihood. Afterwards, we presented three concrete examples of inverse problems that used heavy-tailed or ID prior measures. In particular, we studied the well-posedness of a deconvolution problem with a Lévy process prior that was cast on the non-separable space  $BV(\mathbb{T})$ .

The results of this article open the door for the use of large classes of prior measures in inverse problems and it can be extended in several directions. For example, we showed that

if the forward problem is linear and the measurement noise is Gaussian then one can achieve well-posedness for priors that have poor tail behavior. Then many of the common heavy-tailed priors can be used to model sparsity in the linear case. But it is not clear which prior is the optimal choice and in what sense it is optimal. Furthermore, given that the Compressed Sensing literature is mainly focused on recovery of sparse signals from linear measurements, it is interesting to study the implications of the Compressed Sensing theory in the setting of Bayesian inverse problems. Throughout the article we mentioned the issue of sparsity on several occasions but this is not the only setting where non-Gaussian priors can be useful. For example, non-Gaussian priors can be used in modelling of constraints or in construction of hierarchical models. Then using the theory that was developed here we can study new classes of hierarchical priors and decide which ones will result in a well-posed inverse problem.

**Acknowledgements.** The author owes a debt of gratitude to Prof. Nilima Nigam for her help in the writing of this article and many useful discussions and comments.

## REFERENCES

- [1] C. D. Aliprantis and K. Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, New York, third edition, 2006.
- [2] D. Applebaum. *Lévy processes and stochastic calculus*. Number 93 in Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, 2009.
- [3] J. M. Bernardo and A. F. Smith. *Bayesian theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2009.
- [4] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, three edition, 2008.
- [5] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, 1998.
- [6] V. I. Bogachev. *Measure theory*, volume 1. Springer, New York, 2007.
- [7] V. I. Bogachev. *Measure theory*, volume 2. Springer, New York, 2007.
- [8] L. Bondesson. A general result on infinite divisibility. *The Annals of Probability*, pages 965–979, 1979.
- [9] C. Borell. Convex measures on locally convex spaces. *Arkiv för Matematik*, 12(1):239–252, 1974.
- [10] G. Buttazzo, M. Giaquinta, and S. Hildebrandt. *One-dimensional variational problems: an introduction*. Number 15 in Oxford Lecture Series in Mathematics and Its Applications. Oxford University Press, Oxford, 1998.
- [11] D. Calvetti, J. P. Kaipio, and E. Somersalo. Inverse problems in the Bayesian framework. *Inverse Problems*, 30(11):110301, 2014.
- [12] D. Calvetti and E. Somersalo. *An introduction to Bayesian scientific computing: Ten lectures on subjective computing*, volume 2 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer Science & Business Media, New York, 2007.
- [13] E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.
- [14] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.
- [15] I. Castillo, J. Schmidt-Hieber, and A. Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.
- [16] I. Castillo and A. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.
- [17] R. Cont and P. Tankov. *Financial modelling with jump processes*. Chapman & Hall/CRC Financial mathematics series. CRC press LLC, New York, 2004.

- 
- [18] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian inverse problems for PDEs. *SIAM Journal on Numerical Analysis*, 48(1):322–345, 2010.
  - [19] M. Dashti, S. Harris, and A. M. Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, 2012.
  - [20] I. Daubechies et al. *Ten lectures on wavelets*. Number 61 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1992.
  - [21] C. Fienup and J. Dainty. Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, pages 231–275, 1987.
  - [22] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Springer Science & Business Media, New York, 2013.
  - [23] P. Ghosh and A. Chakrabarti. Posterior concentration properties of a general class of shrinkage priors around nearly black vectors. *arXiv preprint at arXiv:1412.8161*, 2014.
  - [24] P. C. Hansen, J. G. Nagy, and D. P. O’leary. *Deblurring images: matrices, spectra, and filtering*. SIAM, Philadelphia, 2006.
  - [25] R. W. Harrison. Phase problem in crystallography. *JOSA A*, 10(5):1046–1055, 1993.
  - [26] C. Heil. *A basis theory primer: Expanded edition*. Applied and Numerical Harmonic Analysis. Springer Science & Business Media, New York, 2010.
  - [27] B. Hosseini, C. Mougenot, S. Pichardo, E. Constanciel, J. M. Drake, and J. M. Stockie. A Bayesian approach for energy-based estimation of acoustic aberrations in high intensity focused ultrasound treatment. *arXiv preprint arXiv:1602.08080*, 2016.
  - [28] B. Hosseini and N. Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. 2016. *arXiv preprint at arxiv:1604.02575*.
  - [29] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, Volume 1: Models and Applications*. John Wiley & Sons, New York, second edition, 2002.
  - [30] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer Science & Business Media, New York, 2005.
  - [31] S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, New York, 2012.
  - [32] V. Kruglov. A note on infinitely divisible distributions. *Theory of Probability & Its Applications*, 15(2):319–324, 1970.
  - [33] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537, 2004.
  - [34] W. Linde. *Probability in Banach spaces: Stable and infinitely divisible distributions*. John Wiley & Sons, New York, 1986.
  - [35] F. Lucka. *Bayesian inversion in biomedical imaging*. PhD thesis, University of Muenster, december 2014.
  - [36] M. Markkanen, L. Roininen, J. M. Huttunen, and S. Lasanen. Cauchy difference priors for edge-preserving Bayesian inversion with an application to X-ray tomography. 2016. *arXiv preprint at arXiv:1603.06135*.
  - [37] S. Nadarajah. The kotz-type distribution with applications. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(4):341–358, 2003.
  - [38] S. Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005.
  - [39] S. Peszat and J. Zabczyk. *Stochastic partial differential equations with Lévy noise: An evolution equation approach*, volume 113 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2007.
  - [40] N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.
  - [41] K.-i. Sato. *Lévy processes and infinitely divisible distributions*. Number 68 in Cambridge studies in advanced mathematics. Cambridge university press, Cambridge, 1999.
  - [42] F. W. Steutel and K. Van Harn. *Infinite divisibility of probability distributions on the real line*. Pure and Applied Mathematics. Marcel Dekker Inc., New York, 2003.
  - [43] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.
  - [44] T. J. Sullivan. Well-posed bayesian inverse problems and heavy-tailed stable Banach space priors. 2016. *arXiv preprint at arxiv:1605.05898*.

- [45] M. E. Taylor. *Partial Differential Equations I: Basic Theory*, volume 115 of *Applied Mathematical Sciences*. Springer Science & Business Media, New York, second edition, 2011.
- [46] M. Unser and P. Tafti. *An introduction to sparse stochastic processes*. Cambridge University Press, Cambridge, 2013.
- [47] M. Unser, P. Tafti, A. Amini, and H. Kirshner. A unified formulation of Gaussian vs. sparse stochastic processes. Part II: Discrete-domain theory. *IEEE Transactions on Information Theory*, 60:3036–3051, 2011.
- [48] M. Unser, P. Tafti, and Q. Sun. A unified formulation of Gaussian vs. sparse stochastic processes. Part I: Continuous-domain theory. *IEEE Transactions on Information Theory*, 60:1945–1962, 2011.
- [49] C. R. Vogel. *Computational methods for inverse problems*. SIAM, Philadelphia, 2002.